Audience and Content Pattern

Aulia Amirullah Zulkarneidi - 220606343

November 15th, 2022

Future Learn Analysis

In this analysis, we carried out an exploratory analysis on 7-year period datasets of a massive open online course (MOOC) developed by Newcastle University and run by the online skills provider FutureLearn for 7 year period. This is a statistical analysis of a course named "Cyber Security: Safety At Home, Online and in Life" which took 3 weeks to complete.

Here we provide the documentation of the process we went through from scratch for any purposes such as future development and understanding of what we did to perform the objectives and goals of the project.

A. Business Understanding

We performed business understanding from our stakeholders to collect what they need and want so that we can determine our final goals of the project and satisfy the stakeholders with the project we've created. The details can be described as follows:

A.1. Background Online courses are widely provided on the internet either with free access or with membership. Customers can easily get access to any courses they want to watch out of curiosity. Many providers do free access to attract new customers to join membership or buy online courses they provide. The audience target might vary across regions/ continents depending on marketing or customers' interests. The pattern of audiences and the total of watchers might also have a fluctuation over the years. It might be a bit taxing to see a pattern manually by an analyst and make a decision about what region should be focused on more and what content needs to be improved as well as the relationship between the contents and the audience across regions.

However, due to the proliferation of information systems and technology, businesses increasingly have the capability to accumulate huge amounts of customer data in large databases [1]. The availability of large volume of data on customers has created opportunities and challenges for businesses including FutureLearn to leverage the data and gain competitive advantage such as extracting insights from data to make a sensible decision.

In order to make a proper decision to satisfy business goals, we may want to consider extracting insights from data to find a pattern. The abundance of information we have might overwhelm us to perform analysis so that we may find a pattern and association from the available data. Given the fact the abundance of data we have, Data Mining can solve our problem.

Data Mining or DM can be defined as the process through which beneficial information is extracted from big datasets using a combination of statistical as well as artificial intelligence methods [2]. Data Mining has an ability to discover associations, patterns and stochastic structures from the available data. Addressing our analytical problems such as to find regions with the most users FutureLearner has, how the pattern changes over 7-year period and the pattern of content and audiences over 7-year period can be solved by using exploratory analysis in Data Mining.

A.2. Business Objectives The business objective by carrying out our analysis is to increase the number of audience in the future. It's essential to modify the contents so that we can attract more audience and complete all the contents.

A.3. Inventory of Resources In this analysis, we were supposed to carry out 7 datasets of Video Statistics. Nonetheless, the datasets for the first and second years of observation were not available. Therefore, we performed analysis by using 5 remaining available datasets. We performed the analysis using R language and RStudio as the tool to perform R codes for our analysis. In terms of computing resources, we used Macbook Pro with processor Intel Core i5 with macOS Big Sur Version 11.7.

A.4. Requirements, Assumptions and Constraints There were some requirements we needed to gain to perform the analysis. Here we break down in the table schedule of completion below:

No.	Task	Requirements	Constraints	Week 1	Week 2	Week 3
1.	Address Business and Data Understanding	Necessary FutureLearn Datasets	The availability of datasets for the first and second years of observation	Y		
2.	Data Preparation, Modeling and Evaluation	-	-	Y	Y	Y
3.	Documentation	-	-			Y

A.5. Risks and Contingencies There might be less risks we have except if the legal of the collection of the datasets are compromised. We say less because the analysis is to identify patterns so that we might know what content we need to provide to attract more audiences and which region we need to focus on more in the future.

There are some actions we might be able to take in order to minimize the risks we have such as the security of our datasets. It's essential to keep the datasets secure especially during deployment so that the users in the FutureLearn might not issue their confidentiality against the provider.

A.6. Terminology Here are some terms we might come across during the report and talking about the analysis:

1. Data wrangling: also referred as data munging defined as the process of transforming and mapping data from one "raw" data form into another format so that it becomes more appropriate and valuable for a variety of downstream purposes such as analytics.
2. Mean: a calculated central value of a set of numbers
3. Pipeline: the process and tools used to gather rather raw data from multiple sources, analyze it, and presents the results in an understandable format.
4. Regression analysis: a form of predictive modelling technique which investigate the relationship between a dependent (target) and independent variable(s) predictor.
5. Business Intelligence (BI): a technique to learn how to effectively use to generate reports and find important trends. It's more descriptive than predictive.

A.7. Costs and Benefits This analysis doesn't cost any pounds at all. However, it has a benefit of customer increase in the future by identifying the regions and content which might need to be improved in the future so that Newcastle University can provide more interactive and fascinating module thorugh FutureLearn.

A.8. Data Mining Objectives Here are the data mining objectives as described as follows:

- Identifying which continent the audience come from who access the most of the contents of "Cyber Security: Safety At Home, Online and In Life";
- Identifying how the pattern of audience distribution changes over 7-year period;
- Identifying the correlation pattern between all contents of the online module with the audiences;

A.9. Project Plan Here is our detailed project plan described as follows:

No.	Stages	Resources	Inputs	Outputs	Duration	Risks	Actions In case of Risks
1.	Business Understanding	Future Learn Website	1. The needs 2. The cost	1. The goals 2. The benefits	Three days	Website is down	The coursework slides
2.	Data Understanding	Datasets	Datasets	Basic Understanding of the correlation of the datasets	Four days	The datasets are unavailable	Email the provider
3.	Data Preparation	Datasets	7 year period datasets of Video Statistics	One combined dataset	Three days	Computer is broken	• Get access to Newcastle University's computers • Make a backup of the project and datasets • Github
4.	Modelling	• The combined dataset • Project Template	• The combined dataset	• Plots and charts	Four days	Computer is broken	• Github

No.	Stages	Resources	Inputs	Outputs	Duration	Risks	Actions In case of Risks
5.	Evaluation and Repetitions	- Datasets - The combined dataset - Project Template	- Datasets - The combined dataset	- One combined dataset - Plots and charts	Four days	- The datasets are unavailable - Computer is broken	- Get access to Newcastle University's computers - Make a backup of the project and datasets - Github
6.	Deployment	-	-	-	-	-	-
7.	Documentation	-	- Project Template - R Mark Down	Pdf	Three days	-	-

A.10. Initial Assessment of Tools and Techniques The tools we need to prepare to set up a project is RStudio and R installed on our local environment. We will also perform some exploratory analysis such as prdocuing numerical and graphical summaries of our datasets.

B. Data Understanding

B.1. Background of Data To perform our analysis, we're provided with 7-year period datasets of a massive open online course (MOOC) developed by Newcastle University and run by the online skills provider FutureLearn for 7 year period. This datasets are of a course named "Cyber Security: Safety At Home, Online and in Life" which took 3 weeks to complete. Nonetheless, some datasets are missing such as 2 first-year datasets of video statistics.

B.2. Data Description Here is some data description we might need so that we can understand what to do during data preparation:

```r
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"
```

Based on our initial data analysis, we might only be interested in step position, title, and region views percentage. Due to the unavailability of the first and second year datasets of video statistics, we only need 5 datasets of video statistics and ignore other datasets for our analysis.

B.3. Data Exploration There were few things we need to explore and discover for our analysis. In this scenario, we will cover the patterns of our audience distribution and the correlation of the contents and audience across continents, exploratory analysis methods. We would also need to mutate the datasets to calculate the number of people so that the number is more obvious than the percentage for our sales team later in the future. The mutation will be carried out during the data preprocessing. We also needed to consider to select an appropriate approach to address the null <NA> values either between removing the rows or with other appropariate methods which would not ruin our analysis.

Our data mining goals as addressed before are to produce a pie chart to see the distribution of the overall of audiences, to produce a bar plot to see how the distribution might change over 7-year period and the line chart to see the comparison between continents against the completion of all the contents and see from which point the audience might lose their interest.

B.4. Data Quality As we saw from the data, it turns out that our data doesn't have null <NA> values. Therefore, we can tell that our data is complete and accurate.

```
is.na(cyber.security.3_video.stats)
is.na(cyber.security.4_video.stats)
is.na(cyber.security.5_video.stats)
is.na(cyber.security.6_video.stats)
is.na(cyber.security.7_video.stats)
```

We performed the null value checking to address the missing values to make ourselves more determined. However, we perform the mutation by changing null values to the mean value of its column for the purpose of reproducibility (just in case of the missing values for a new dataset).

C. Data Preparation

C.1. Data Selection As covered before, we might only be interested in step position, title, and continent views percentage from 5-year period datasets of video statistics. As stated above, we don't have an issue of data quality as it's complete and accurate. However, we also consider data quality issue which might rise in the future by checking the null values in our pipeline.

C.2. Data Cleaning Even though, we don't have an issue of data quality, we might want to address data cleaning for the purpose of reproducibility of our project. Since our data is statistics value, we might not want to remove the rows as it might ruin our analysis for either the pattern of audience distribution and the pattern of our content distribution associated with audience. There are varieties of algorithm which are suggested to handle missing data , with different degree of complexity: trimmed score method (TRI), single-component projection (SCP), projection to the model plane (PMP) – using PLS or ordinary least squares (PMP_{PLS}, PMP_{OLS}), iterative imputation of missing data (II), a method based on the minimization of the squared prediction error (SPE), conditional mean replacement (CMR), trimmed score regression (TSR), and regression on known data (KDR) [3]. Therefore, we might only consider the mean replacement as it's relatively relevant to address the number of audience if we might have missing values based on the pattern of the mean value of each column of the continent audience percentage,

C.3. Data Construction In this stage, we performed a mutation to create new columns such as year and the numbers of audience across continents. The way we calculated the total number of audience is by dividing the percentage with 100 and multiplied it with the total views in each step position. Hence, we get new brand columns of continents views percentage. During the construction, we also include step position, title and total views for the reproducibility and for the purpose of our plots and charts we will produce during modelling.

C.4. Data Integration During data integration, we integrate all the variables created during the previous stages. In our scenario, we only have 5 datasets to be considered. Therefore, we piled up those 5 datasets which have the same number of columns with the same order and name of columns on top of each other using rbind.

```
# combine all preprocessed data into a brand new data frame
datalist = list() ## make an empty list
for(i in 1:length(datasets)) {
  if (!is.null(datasets[i][[1]])) {
    datalist[i] <- lapply(paste0(variable_name, i), get)  # get variables
  }
}

res = do.call(rbind, datalist) ## stack datasets into one dataset
```

We instantiated our datalist to accumulate all existant variables. Then, we looped through our predefined datasets and only put variables which exist. After all the process, we row bind all the data and put it into the final variable called res.

C.5. Data Format We may format our data and convert it into a csv. However, since we use projectTemplate, we don't have to format our data and we can carry on with our newly created variabled called res which contain all the information we need to perform modelling.

D. Modelling

D.1. Modelling Technique In this stage, we will discuss the technique we use to achieve one of the objectives we have. In this modelling for the first cycle which is to answer which continent the audience of our online course gets access to more so that we can focus more on some of those continents for ads, online reviews and course improvements so that we won't lose audience but perhaps we can get more audience later in the future.

There are some techniques we use to produce a pie chart and a bar plot to feed the first cycle. The first technique we use is to build a pipeline which group the data by year and perform mean calculation of each continent views number to get the proportion and feed it into our chart and plot.

```
stats_summaries_number <- res %>%
  group_by(year) %>%
  summarise(europe = mean(europe_views_number),
            oceania = mean(oceania_views_number),
            asia = mean(asia_views_number),
            north_america = mean(north_america_views_number),
            south_america = mean(south_america_views_number),
            africa = mean(africa_views_number),
            antarctica = mean(antarctica_views_number))
knitr::kable(head(stats_summaries_number), "simple")
```

year	europe	oceania	asia	north_america	south_america	africa	antarctica
3	468.9231	22.46154	80.38462	84.61538	18.769231	51.92308	0
4	411.9231	30.07692	135.76923	75.53846	19.384615	112.53846	0
5	485.3077	38.15385	83.38462	120.61538	12.000000	46.92308	0
6	206.4615	20.92308	135.30769	33.23077	10.692308	34.61538	0
7	248.9231	16.84615	71.38462	31.69231	9.846154	23.30769	0

The next step, we did is to transpose the data we have as we want to show the continents as the x values and year as the y values. Therefore transposing is crucial in this stage to feed our objective. We also need to convert it into a dataframe.

```
df2 <- data.frame(t(stats_summaries_number[-1]))
knitr::kable(head(df2), "simple")
```

	X1	X2	X3	X4	X5
europe	468.92308	411.92308	485.30769	206.46154	248.923077
oceania	22.46154	30.07692	38.15385	20.92308	16.846154
asia	80.38462	135.76923	83.38462	135.30769	71.384615
north_america	84.61538	75.53846	120.61538	33.23077	31.692308
south_america	18.76923	19.38462	12.00000	10.69231	9.846154
africa	51.92308	112.53846	46.92308	34.61538	23.307692

Then, we also want the column name of our continent in the dataframe so that we can address it during the plottings. We also perform the deletion of our unnamed column which is actually our continents.

```
## Copy the first column to create a new column called continent
df2 <- data.frame("continents"= rownames(df2), df2)

## Delete columns which don't have column title
row.names(df2) <- NULL
colnames(df2)
```

```
## [1] "continents" "X1"         "X2"         "X3"         "X4"        
## [6] "X5"
```

Next, since we want to display all the values in x axis and leave the values of the number of audience on the y axis, we will append all the values of the number of audience across continents on top of each other. The algorithm we use should allow us to reshape and elongate the data frames in a defined-manner so that it organizes the data values in a long data frame format. We use melt algorithm in this case to reshape our data frame.

```r
# Melt data frame into long format
df3 <- melt(df2 ,  id.vars = 'continents', variable.name = 'year')
knitr::kable(head(df3), "simple")
```

continents	year	value
europe	X1	468.92308
oceania	X1	22.46154
asia	X1	80.38462
north_america	X1	84.61538
south_america	X1	18.76923
africa	X1	51.92308

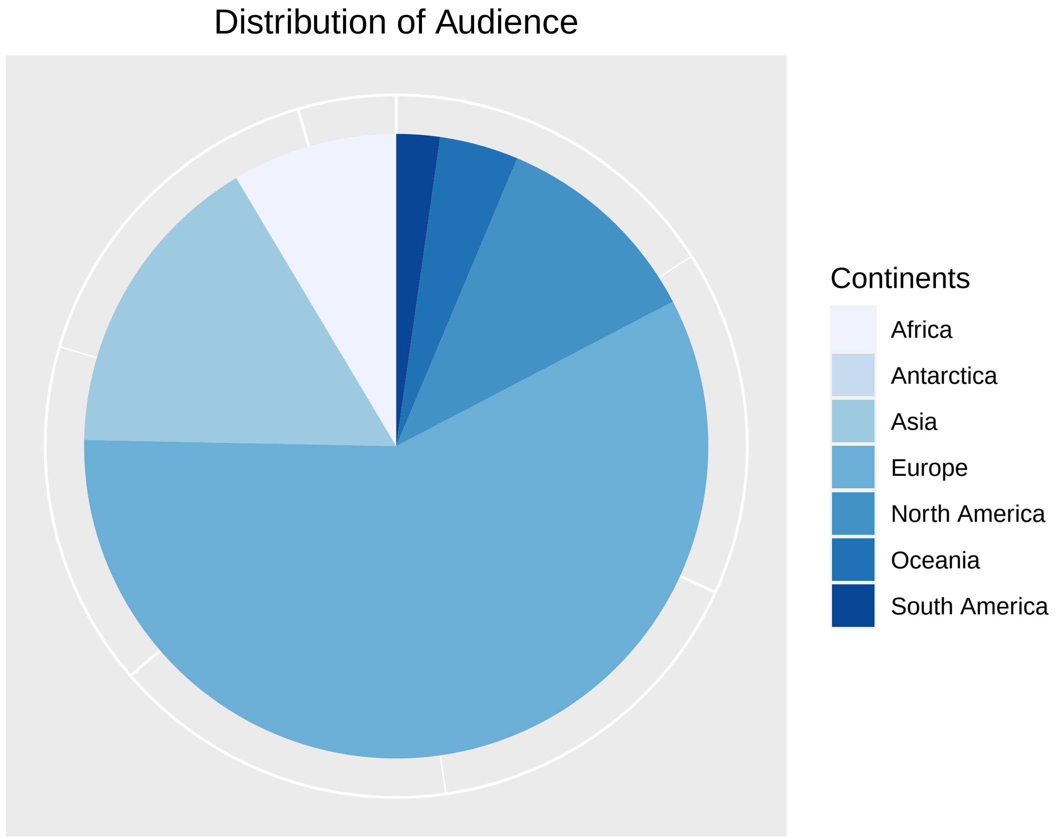
We also need to combine all the continents to see the proportion as a whole. Therefore, we will perform another pipeline by grouping all the continents together and summarize the value as follows:

```r
## Combine the value for all years into one to fit into our pie chart
## to get an idea of the distribution of our audience
df_joined <- df3 %>% group_by(continents) %>% summarise(value=mean(value))
knitr::kable(head(df_joined), "simple")
```

continents	value
africa	53.86154
antarctica	0.00000
asia	101.24615
europe	364.30769
north_america	69.13846
oceania	25.69231

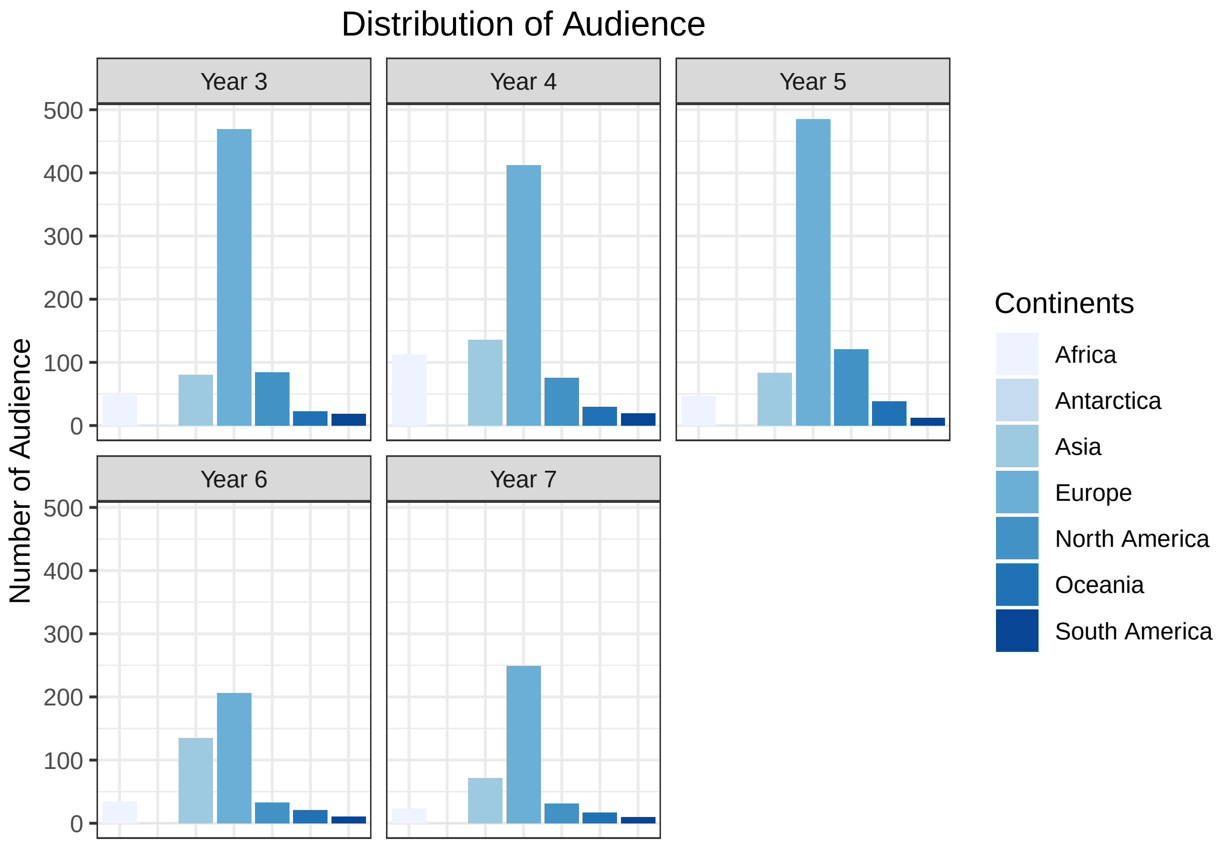
We can perform the plottings now as we've got the dataframe we want to feed the first cycle.

```r
# Barplot and Pie Chart of Our Overall Distribution
bp <- ggplot(df_joined, aes(x="", y=value, fill=continents))+
  geom_bar(width = 1, stat = "identity")

pie <- bp + coord_polar("y", start=0) +
  labs(title = "Distribution of Audience") +
  theme(axis.title=element_blank(),
        axis.text=element_blank(),
        axis.ticks=element_blank(),
        plot.title=element_text(hjust=0.5)) +
  scale_fill_brewer("Continents",
                    labels = c("Africa", "Antarctica", "Asia",
                               "Europe", "North America",
                               "Oceania", "South America"),
                    palette="Blues")
pie
```



As we can tell, for the overall distribution, many of our audiences come from europe with asia as the second most audience and North America as the third. We may target three of those continents for ads more in the future at glimpse. However, we need to see how the break down might look like. Therefore, we will try to perform a bar plot to see how the proportion and how it changes over 7-year period.

```
# New facet label names for year variable
year.labs <- c("Year 3", "Year 4", "Year 5", "Year 6", "Year 7")
names(year.labs) <- c("X1", "X2", "X3", "X4", "X5")

# Create line plot for each column in data frame
ggplot(df3, aes(x = continents, y = value, fill = continents)) +
  geom_bar(stat = 'identity') +
  facet_wrap(~ year,
             labeller = labeller(year = year.labs),
             nrow=2)  +
  labs(title = "Distribution of Audience",
       x = "Continents",
       y = "Number of Audience",
       color = "Continents") +
  theme_bw() +
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        plot.title=element_text(hjust=0.5)) +
  scale_fill_brewer("Continents",
                    labels = c("Africa", "Antarctica", "Asia",
                               "Europe", "North America",
                               "Oceania", "South America"),
                    palette="Blues")
```



As we can see from the bar plots, There's a decline after year 5 for our audience in Europe to watch our online course. Even though, there's a decrease in number of our audience in Europe since the fifth year, our audience mostly comes from Europe with Asia comes the second most of the years and North America comes the third for the most of our years observation.

To feed the second cycle, we use our original data to see the proportion our audience from those continents we described in the first cycle with the contents so that we see the completeness pattern across regions so that we can determine from which content, we start to lose audience.

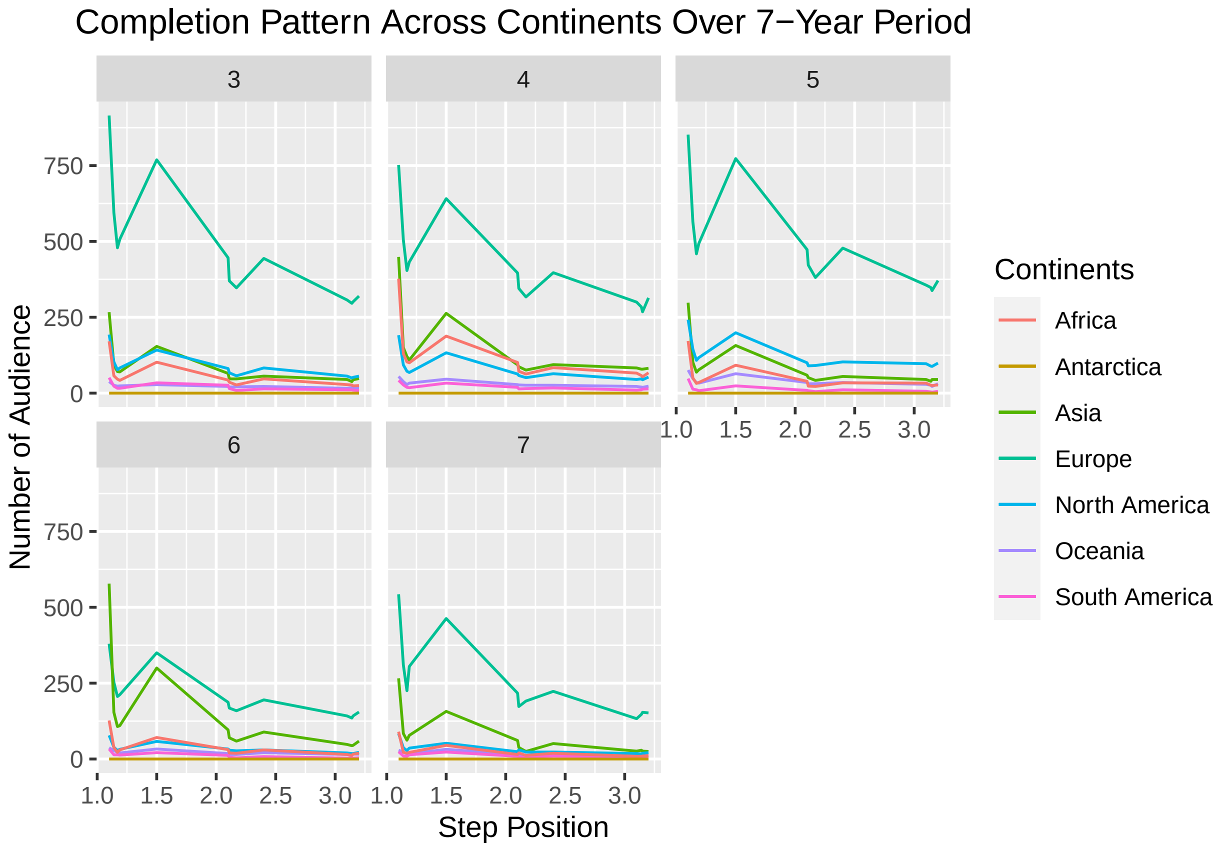
We come back to business understanding to feed the second cycle. As the objective of our business project is to increase the number of audience in the future, we need to see the correlation between the contents and the audience for the second cycle so that the module leaders might want to take an action on how to improve the attractiveness or modification on the contents they might want to deliver in order to attract more audience and complete all the contents. To feed the second cycle, the inventory of resources, we have is the same with what we needed to feed the first cycle such as R language and RStudio as the tool to perform R codes for our analysis. In terms of computing resources, we used Macbook Pro with processor Intel Core i5 with macOS Big Sur Version 11.7. We don't need to change our computing resources as our device still can handle the modelling as the data is not quite big during the development.

In the first cycle, we've identified which continents the audience come from who access the most of the contents of "Cyber Security: Safety At Home, Online and In Life" and the distribution pattern of audience changes over 7-year period. In the second cycle, our data mining objectives will shift to Identifying the correlation pattern between all contents of the online module with the audiences. The first cycle is correlated to the second as we've seen which continents whom our module gets the most of the audience from. So, we can focus on those regions for ads and fix the contents later in the future.

In terms of data understanding, we use the same datasets. Therefore, no difference we might get either. Our data quality is already addressed as complete. Our data preparation is also well addressed as in the first cycle. We use the same data preprocessing process as same as the first cycle fed. We're only interested in step position, title, and continent views percentage from 5-year period datasets of video statistics. However, we also consider data quality issue which might rise in the future by checking the null values in our pipeline. We use the same technique as the first cycle did such as the mean replacement to clean the null values.

Now, we need to come back to the modelling by performing a line chart so that we know the pattern as follows:

```
ggplot(res, aes(step_position)) +
  geom_line(aes(y = europe_views_number, colour = "Europe")) +
  geom_line(aes(y = oceania_views_number, colour = "Oceania")) +
  geom_line(aes(y = asia_views_number, colour = "Asia")) +
  geom_line(aes(y = north_america_views_number, colour = "North America")) +
  geom_line(aes(y = south_america_views_number, colour = "South America")) +
  geom_line(aes(y = africa_views_number, colour = "Africa")) +
  geom_line(aes(y = antarctica_views_number, colour = "Antarctica")) +
  facet_wrap(~ year) +
  labs(title = "Completion Pattern Across Continents Over 7-Year Period",
       x = "Step Position",
       y = "Number of Audience",
       color = "Continents") +
  theme(plot.title=element_text(hjust=0.5))
```



From oursecond cycle, we can tell that there is a pattern of the distribution of our audience in all regions against the content that they're interested in to explore. In the beginning of the course, we have lots of people getting started but suddenly we didn't have audience to watch the rest of the contents except in the module privacy online and offline before it started to drop again in the next episodes. It seems that in many regions, they're just mostly interested in that episode and lose interest from that episode.

D.2. Test Design During this stage, we don't really perform any test design either describing how the models are built, tested or evaluated as the goal of the project is to identify a pattern which might help the company to make a better module and focus more on certain regions for ads.

D.3. Model Assessment In this stage, we might be able to assess our model based on its accuracy or performance. However, our model is an exploratory analysis. Therefore, we could only tell that the produced models for the first and second cycle are legitimate and accurate based on the datasets provided.

E. Evaluation

E.1. Results Evaluation We've already evaluated the results which are the the continents who access the most of our contents and how it changes over 7-year period and for the second cycle, we've seen that most of our audience in all continents lost interest to carry on with our videos since the "privacy online and offline" module. Our models are considered valid and met the project objectives and satisfied the data mining goals too.

E.2. Process Review Our process which might want to be highlighted is the producibility. During the data preprocessing in the data preparation section of CRISP-DM model, we put our wanted datalist in a datasets variable. The aim is to address the producibility so that if there's another dataset, we can only put it in the list variable called dataset.

```
# List all our datasets.
datasets = list(NULL, NULL, cyber.security.3_video.stats, cyber.security.4_video.stats,
                cyber.security.5_video.stats, cyber.security.6_video.stats,
                cyber.security.7_video.stats)
```

We also create a variable which will convert a string into a variable so that our data preprocessing will be more automatic without creating more variables manually when there's another new dataset put into our analysis.

```
variable_name = "video_stats_"
```

In terms of the producibility during data cleaning and wrangling (munging), we loop through the whole list of our datasets so that our dataset variable can be automatically created. We're doing this because we will put all the newly automatedly created dataset variables together on top of each other.

```
# DATA WRANGLING
for(i in 1:length(datasets)) { # loop through the datasets list
  if (!is.null(datasets[i][[1]])) { # Datasets of the year 1 and 2 are not available
    temp <- datasets[i][[1]] %>%
      select(
        step_position:total_views, # select all
        #columns from step_position to total_views from the raw data
        ends_with("views_percentage")) %>% # select all columns
      # with the ending "views_percentage"
      mutate(europe_views_percentage = replace_na(europe_views_percentage,
                                                  mean(europe_views_percentage)),
             oceania_views_percentage = replace_na(oceania_views_percentage,
                                                   mean(oceania_views_percentage)),
             asia_views_percentage = replace_na(asia_views_percentage,
                                                mean(asia_views_percentage)),
             north_america_views_percentage = replace_na(north_america_views_percentage,
                                                         mean(north_america_views_percentage)),
             south_america_views_percentage = replace_na(south_america_views_percentage,
                                                         mean(south_america_views_percentage)),
             africa_views_percentage = replace_na(africa_views_percentage,
                                                  mean(africa_views_percentage)),
             antarctica_views_percentage = replace_na(antarctica_views_percentage,
                                                      mean(antarctica_views_percentage))) %>%
      mutate(year = i, # add a column called year
             europe_views_number = round((europe_views_percentage / 100)
                                         * total_views, digits = 0),
             # add columns by converting percentage to number
             oceania_views_number = round((oceania_views_percentage/ 100)
                                          * total_views, digits = 0),
             asia_views_number = round((asia_views_percentage / 100)
                                       * total_views, digits = 0),
             north_america_views_number = round((north_america_views_percentage
                                                 / 100) * total_views,
                                                digits = 0),
             south_america_views_number = round((south_america_views_percentage
                                                 / 100) * total_views,
                                                digits = 0),
             africa_views_number = round((africa_views_percentage / 100) *
                                           total_views, digits = 0),
             antarctica_views_number = round((antarctica_views_percentage /
                                                100) * total_views, digits = 0))

    # create a new variable by using assign and paste0 to combine strings
    assign(paste0(variable_name,i), temp)
  }
}
```

The command “assign” will create a new variable based on our dataset which is looped through in the data selection, cleaning and wrangling process. Then, we combine all the variables together on top of each other.

```
# combine all preprocessed data into a brand new data frame
datalist = list() ## make an empty list
for(i in 1:length(datasets)) {
  if (!is.null(datasets[i][[1]])) {
    datalist[i] <- lapply(paste0(variable_name, i), get)  # get variables
  }
}

res = do.call(rbind, datalist) ## stack datasets into one dataset
```

Our data preprocessing use the concept of reproducibility to address many scenarios such as a new dataset is added into our datasets and a new data analyst willing to modify our pipeline so that the data addition and modification will not be taxing. Our newly created dataset can also be replicated for sicence projects or other related projects. In this case, we're addressing the replicability of our analysis and the dataset we've just created.

E.3. Next Steps and Actions There are several things we might be interested in to carry on with our analysis. We might want to consider Logistic regression, Neural Networks or Decision Tree to predict future numbers of audience in the future or implement other algorithms to carry out for any predictions related to contents and audience.

F. Deployment

F.1. Deployment Plan We don't necessarily need to plan deployment scheme. However, for the sake of the future, we may consider to deploy it on the IBM clouds using Software as A Service (SaaS). We'll take advantage of cloud computing infrastructure and economies of scale. We might need to open a new account and use Pay as You Go.

F.2. Monitoring and Maintenance Plan We don't really need to perform a scheme to monitor and maintain our models as it's an exploratory analysis.

References

[1] Michael J Shaw, Chandrasekar Subramaniam, Gek Woo Tan, Michael E Welge, Knowledge management and data mining for marketing, Decision Support Systems, Volume 31, Issue 1, 2001, Pages 127-137, ISSN 0167-9236, https://doi.org/10.1016/S0167-9236(00)00123-8

[2] Mehrbakhsh Nilashi, Behrouz Minaei-Bidgoli, Abdullah Alghamdi, Mesfer Alrizq, Omar Alghamdi, Fatima Khan Nayer, Nojood O Aljehane, Arash Khosravi, Saidatulakmal Mohd, Knowledge discovery for course choice decision in Massive Open Online Courses using machine learning approaches, Expert Systems with Applications, Volume 199, 2022, 117092, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.117092.

[3] Steven D. Brown, Romá Tauler and Beata Walczak, Comprehensive Chemometrics (Chemical and Biochemical Data Analysis), Volume 4, 2009, ISSN 978-0-444-52701-1, Poland.