CSC8631 Assignment Report

D Walmsley, C1053068

15/11/2021

Introduction

The purpose of this report is to analyse data collected from an online course ran by Futurelearn (www.futurelearn.com). The course is ran in collaboration with Newcastle University. The report will demonstrate findings from a data set across 7 runs provided by the the university. Using the CRISP-DM methodology to provide a clear structure to the analysis. This report will be broken into sections that are the key steps of the CRISP DM process. The sections are *Business Understanding, Data Understanding, Data preparation, Modelling, Evaluation and Development.*

I.Business Understanding

Newcastle University is a Russell group university offering high quality courses to students from around the world. Working with online partner Future Learn, the university are offering an online cyber security course. The course provider must show awareness of factors impacting on cohort recruitment such as demographic, age, and gender. The provider must review key performance indicators and analyse how appealing the course is to specific user groups. Course retention and understanding of students motive to enroll must be evaluated. In addition the provider must review barriers and challenges that sub groups of the cohort have experienced and how this may have prevented them completing the course. This will allow the provider to address issues in future and develop a sustainable product. Academic outcomes of students are of interest to the course providers as this allows them to identify what resources are having the greatest impact on learning, but also allows them to identify areas where improved resources would impact on student outcomes. Finally the university must consider how it could keep up with changes in new technology and how students access the content provided. Accessibility is a key factor and how the provider can ensure a course or product is more accessible to a wider audience, encouraging a greater enrolment globally.

The data files from 7 runs or repetitions of the course have been acquired. These files will be analysed to determine key links and trends within the data. The strength of the data collected will determine the strength of potential outcomes of this project. The data from the repetitions or reps needs to have similarities to ensure that files and scenarios are comparable. Clear identifiable relationships between the data files will help construct stronger lines of enquiry. This will lead to outputs that will have a greater impact on informing future decision making within the process. From the third course rep information is included about video resources used within the course material and how and where the videos were accessed. Although this video data was not accessible for reps one and two, this may be a way to compare if video files have had an impact since their introduction in rep three of the course. R scripts will be written within R studio software package to ensure that any work undertaken can be easily reproduced.

This project will look to meet the aims and goals set out in the business objectives. By exploring the data, the intention of this project is to identify possible links between the introduction of video resources and the completion of the course. Is it possible to identify a groups of students who are more likely to complete the course and is there a clear reason for this? Are there clear indications why certain students fully complete the course or actually drop out of the course? By using the data to identify clear strengths and weaknesses of the course, this should inform decision making on what resources need strengthening or where to better support students in order to improve retention and academic outcomes.

R scripts will be used in order to clean and collate data. The use of R language allows the data not only cleaned, but also means that the data can be displayed in clear visual formats that support what large datasets are saying. The R scripts will be accessed through R studio software that allows the scripts to be stored within a Project Template structure. This means managing and accessing scripts is both organised and accessible. R studio then allows these scripts, visual data representations and charts to be merged together into a report. This can be exported as a PDF file format that can be easily distributed.A key success criteria of this project is reproducibility and the CrispDM methodology gives a clear structure to the process and reproduction of the work taken place within the project. This methodology will be structured through the Project Template file structure to ensure that other users could replicate the work that had been undertaken.

A Gitlog will be included as part of the project to show version control techniques and ensure that changes to the work are recorded at regular intervals in the event of errors occurring. Version control allows all work to be rolled back to staged points in time.

II. Data Understanding

There are 53 .csv files provided by the university, spread across seven repetitions of the course. The files all start with the prefix cyber.security then the number of the repetition of the course (cyber.security.1, cyber.security.2 etc). These are then split into eight possible datasets in the final five repetitions of the course and only seven possible datasets from the first two repetitions. This is because there is no "videostats" files until repetition 3 onwards. The files provided by the university are as follows:

> Archetype survey responses
> Enrolments
> Leaving survey responses
> Question responses
> Step activity
> Team members
> Video stats
> Weekly sentiment survey responses

Describe data:

Examine the data and document its surface properties like data format, number of records, or field identities.

Unfortunately nine of the data files included no data within the files, this is a weakness in the data collection process. These files were:

- Cyber.security.1_archetype.survey.responses.csv
- Cyber.security.1_leaving.survey.responses.csv
- Cyber.security.1_weekly.sentiment.survey.responses.csv
- Cyber.security.2_archetype.survey.responses.csv
- Cyber.security.2_leaving.survey.responses.csv
- Cyber.security.2_weekly.sentiment.survey.responses.csv
- Cyber.security.3_leaving.survey.responses.csv
- Cyber.security.3_weekly.sentiment.survey.responses.csv
- Cyber.security.4_weekly.sentiment.survey.responses.csv

Looking at the enrolments data, the enrolment files were combined to see the total number of students that had enrolled over the seven repetitions.

```
glimpse(combined_enrolments)
```

```
## Rows: 37,296
## Columns: 14
## $ learner_id                 <chr> "160d6600-ea0e-4568-bfa9-5d7cd5b8e61b", "4dc22~
## $ enrolled_at                <chr> "2016-08-10 14:28:49 UTC", "2016-05-24 17:34:3~
## $ unenrolled_at              <chr> "", "2018-10-30 20:20:51 UTC", "", "", "", "",~
## $ role                       <chr> "learner", "learner", "learner", "learner", "l~
## $ fully_participated_at      <chr> "", "", "2016-09-22 16:56:03 UTC", "", "", "20~
## $ purchased_statement_at     <chr> "", "", "", "", "", "", "", "", "", "", "", ""~
## $ gender                     <chr> "Unknown", "male", "Unknown", "Unknown", "Unkn~
## $ country                    <chr> "Unknown", "PE", "Unknown", "Unknown", "Unknow~
## $ age_range                  <chr> "Unknown", "46-55", "Unknown", "Unknown", "Unk~
## $ highest_education_level    <chr> "Unknown", "university_degree", "Unknown", "Un~
## $ employment_status          <chr> "Unknown", "working_part_time", "Unknown", "Un~
## $ employment_area            <chr> "Unknown", "teaching_and_education", "Unknown"~
## $ detected_country           <chr> "GB", "PE", "NG", "UG", "IM", "NO", "GB", "GB"~
## $ rep                        <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
```

To combine the tables, the following code was used to combine the cyber security files. This script takes the original cyber.security file and renames it with a name that is easier to remember and reuse later in enrolments_yr1 for example, then adds a column that indicates the repetition or run of the course. It then combines the enrolment_yr files into one dataframe that can be used later in the project.

```
enrolments_yr1 = cyber.security.1_enrolments %>%
  mutate(rep = "rep 1")
enrolments_yr2 = cyber.security.2_enrolments %>%
  mutate(rep = "rep 2")
enrolments_yr3 = cyber.security.3_enrolments %>%
  mutate(rep = "rep 3")
enrolments_yr4 = cyber.security.4_enrolments %>%
  mutate(rep = "rep 4")
enrolments_yr5 = cyber.security.5_enrolments %>%
  mutate(rep = "rep 5")
enrolments_yr6 = cyber.security.6_enrolments %>%
  mutate(rep = "rep 6")
enrolments_yr7 = cyber.security.7_enrolments %>%
  mutate(rep = "rep 7")

combined_enrolments = rbind(enrolments_yr1, enrolments_yr2, enrolments_yr3, enrolments_yr4, enrolments_
```

From the enrolments data every learner had a unique learner ID and these learner ids could be found in the archetype, leaving response survey, question response survey, step activity, team members and weekly sentiment survey tables. These relational data tables make combing the files and looking for trends easier and show the individuals learning journey as a whole. Within the enrolment data there is a number of blank cells, as well as cells that contain “Unknown”. There are examples of this within the columns gender, age range, highest education level and employment area.

```
sum(combined_enrolments$gender == "Unknown")
```

```
## [1] 33137
```

```r
sum(combined_enrolments$age_range == "Unknown")
```

```
## [1] 33268
```

```r
sum(combined_enrolments$highest_education_level == "Unknown")
```

```
## [1] 33161
```

```r
sum(combined_enrolments$employment_area == "Unknown")
```

```
## [1] 34090
```

There are also a number of cells that were empty or blank cells. These cells appeared in the purchased statement, unenrolled and fully participated column. T

```r
sum(combined_enrolments$purchased_statement_at == "")
```

```
## [1] 37007
```

```r
sum(combined_enrolments$unenrolled_at == "")
```

```
## [1] 33105
```

```r
sum(combined_enrolments$fully_participated_at == "")
```

```
## [1] 35142
```

The question response data shows each users responses to the staged quizzes throughout the course. It also contains whether their responses were correct and what answers were provided to the question. This data can be linked to other data files using the individual learner ids that are provided. As for the enrolment data files the same process was completed to create a large file to check for missing data. In following this process the data would know have the rep or repetition column added to the end.

```r
glimpse(qresponse_all)
```

```
## Rows: 176,463
## Columns: 11
## $ learner_id      <chr> "77454a73-6b8b-46a2-8dee-35f36b6c4fc1", "77454a73-6b8b~
## $ quiz_question   <chr> "1.7.1", "1.7.1", "1.7.1", "1.7.1", "1.7.1", "1.7.1", ~
## $ question_type   <chr> "MultipleChoice", "MultipleChoice", "MultipleChoice", ~
## $ week_number     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ step_number     <int> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ~
## $ question_number <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ response        <chr> "1,2", "1,2,3", "1,2,3", "1,2", "2,3", "1,2,3", "1,2,3~
## $ cloze_response  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ submitted_at    <chr> "2016-07-06 10:37:05 UTC", "2016-07-06 10:57:05 UTC", ~
## $ correct         <chr> "false", "true", "true", "false", "false", "true", "tr~
## $ rep             <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", ~
```

This data contains a large amount of NA responses in the close response column.

```
  sum(is.na(qresponse_all$cloze_response))
```

```
## [1] 176463
```

By running the code you can see the qresponse_all table contains 176,463 rows of data and in the cloze_response column there is 176,463 columns containing NA.

The introduction of video stats begins in repetition 3 of the course. Within the data files are the title of each video, the video duration, total number of views, the type of device it has been watched on, whether the file was downloaded, the percentage of the video a user has watched as well as the continent where it was watched.

```
glimpse(combined_videostats)
```

```
## Rows: 65
## Columns: 29
## $ step_position                  <dbl> 1.10, 1.14, 1.17, 1.19, 1.50, 2.10, 2.1~
## $ title                          <chr> "Welcome to the course", "Why would any~
## $ video_duration                 <int> 99, 362, 241, 348, 281, 37, 312, 92, 42~
## $ total_views                    <int> 1659, 910, 723, 755, 1248, 694, 564, 51~
## $ total_downloads                <int> 113, 77, 63, 62, 100, 48, 53, 42, 50, 3~
## $ total_caption_views            <int> 36, 8, 5, 2, 15, 1, 4, 3, 5, 1, 1, 5, 4~
## $ total_transcript_views         <int> 221, 173, 120, 147, 191, 108, 110, 87, ~
## $ viewed_hd                      <int> 58, 28, 16, 10, 41, 13, 434, 7, 16, 6, ~
## $ viewed_five_percent            <dbl> 76.97, 72.53, 73.72, 72.85, 78.45, 76.3~
## $ viewed_ten_percent             <dbl> 75.35, 70.88, 73.86, 71.92, 75.64, 75.0~
## $ viewed_twentyfive_percent      <dbl> 73.42, 68.57, 71.92, 69.27, 69.87, 74.9~
## $ viewed_fifty_percent           <dbl> 70.40, 65.38, 69.71, 64.90, 65.63, 73.4~
## $ viewed_seventyfive_percent     <dbl> 68.17, 63.08, 66.11, 63.44, 62.66, 72.9~
## $ viewed_ninetyfive_percent      <dbl> 66.43, 61.54, 61.83, 61.59, 59.05, 71.1~
## $ viewed_onehundred_percent      <dbl> 63.71, 56.81, 44.67, 49.40, 44.87, 69.4~
## $ console_device_percentage      <dbl> 0.06, 0.11, 0.14, 0.13, 0.00, 0.14, 0.1~
## $ desktop_device_percentage      <dbl> 78.60, 79.23, 79.67, 78.54, 80.37, 79.1~
## $ mobile_device_percentage       <dbl> 13.26, 10.33, 8.71, 9.40, 11.38, 9.37, ~
## $ tv_device_percentage           <dbl> 0.06, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0~
## $ tablet_device_percentage       <dbl> 7.72, 10.11, 11.07, 11.39, 7.93, 10.95,~
## $ unknown_device_percentage      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ europe_views_percentage        <dbl> 55.15, 65.38, 66.25, 67.15, 61.62, 64.2~
## $ oceania_views_percentage       <dbl> 2.29, 2.86, 3.18, 3.18, 2.24, 3.17, 3.5~
## $ asia_views_percentage          <dbl> 16.09, 10.22, 9.82, 9.27, 12.34, 9.37, ~
## $ north_america_views_percentage <dbl> 11.63, 11.32, 10.65, 10.99, 11.38, 11.6~
## $ south_america_views_percentage <dbl> 3.07, 2.53, 2.21, 2.12, 2.72, 3.75, 2.6~
## $ africa_views_percentage        <dbl> 10.31, 6.26, 6.36, 5.56, 8.17, 6.20, 6.~
## $ antarctica_views_percentage    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ rep                            <chr> "rep 1", "rep 1", "rep 1", "rep 1", "re~
```

The video stats files contain a large proportion of numerical data that could be further investigated later in the project.

```
sum(combined_videostats == " ")
```

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

There were no empty cells from the 5 reps of video stats

III. Data Preparation

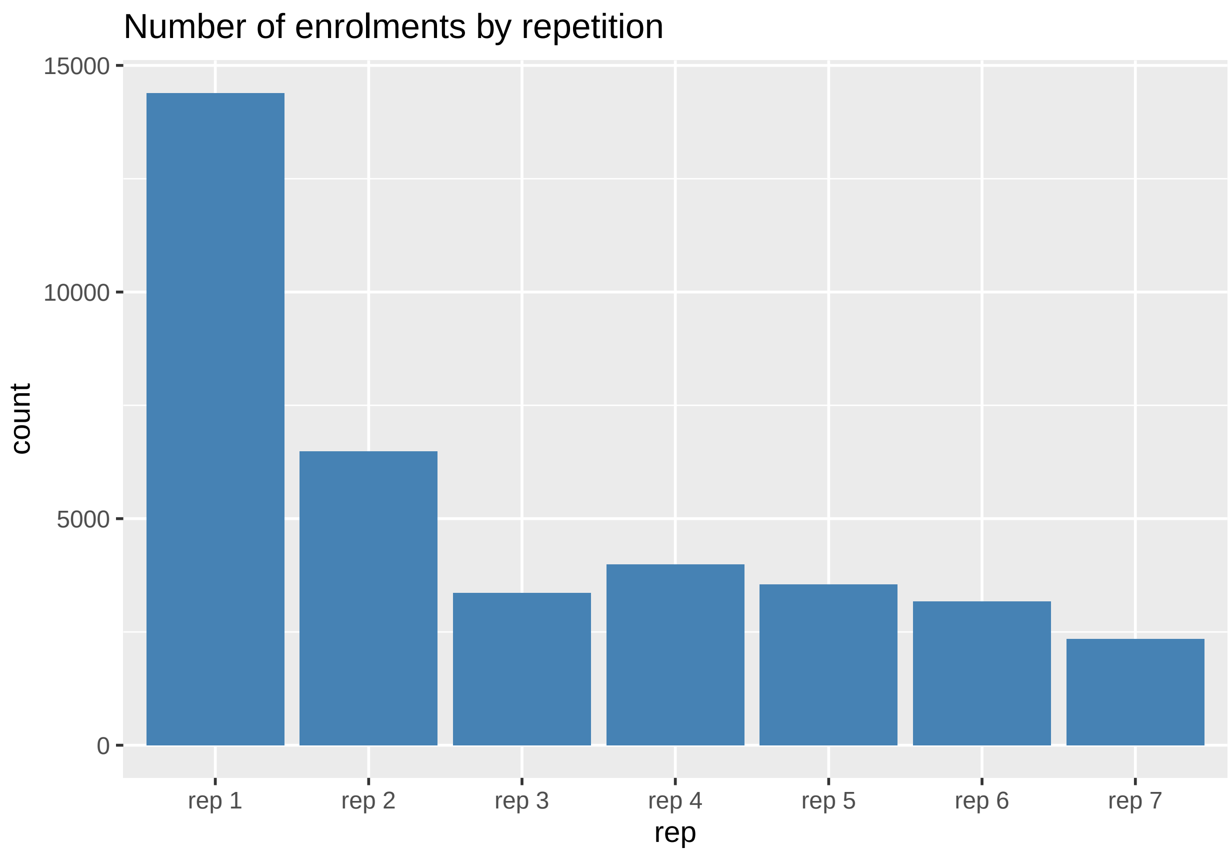
Firstly information was collated on the number of enrolments each year.The enrolments tables has been grouped in the table combined_enrolments.

```
glimpse(combined_enrolments)
```

```
## Rows: 37,296
## Columns: 14
## $ learner_id                 <chr> "160d6600-ea0e-4568-bfa9-5d7cd5b8e61b", "4dc22~
## $ enrolled_at                <chr> "2016-08-10 14:28:49 UTC", "2016-05-24 17:34:3~
## $ unenrolled_at              <chr> "", "2018-10-30 20:20:51 UTC", "", "", "", "",~
## $ role                       <chr> "learner", "learner", "learner", "learner", "l~
## $ fully_participated_at      <chr> "", "", "2016-09-22 16:56:03 UTC", "", "", "20~
## $ purchased_statement_at     <chr> "", "", "", "", "", "", "", "", "", "", "", ""~
## $ gender                     <chr> "Unknown", "male", "Unknown", "Unknown", "Unkn~
## $ country                    <chr> "Unknown", "PE", "Unknown", "Unknown", "Unknow~
## $ age_range                  <chr> "Unknown", "46-55", "Unknown", "Unknown", "Unk~
## $ highest_education_level <chr> "Unknown", "university_degree", "Unknown", "Un~
## $ employment_status          <chr> "Unknown", "working_part_time", "Unknown", "Un~
## $ employment_area            <chr> "Unknown", "teaching_and_education", "Unknown"~
## $ detected_country           <chr> "GB", "PE", "NG", "UG", "IM", "NO", "GB", "GB"~
## $ rep                        <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
```

To identify if there was a trend in the number of enrolments in each year, the number of enrolments over each year is displayed in the below bar chart.

```
ggplot(combined_enrolments, aes(x=rep, fill=rep))+
  geom_bar(fill = "steelblue")+
  theme_grey()+
  ggtitle("Number of enrolments by repetition")
```



Enrolments clearly decline from repetition 1 to 7, however this graph is simply based on the number of learners enrolled on the course at each repetition. It would be helpful to see the number of people who complete the course each repetition. To do this there is a column within the enrolments data which indicates the date a student has fully participated in the course

```
glimpse(combined_enrolments$fully_participated_at)
```

```
##  chr [1:37296] "" "" "2016-09-22 16:56:03 UTC" "" "" ...
```

This column does however contain blank cells.

```
sum(combined_enrolments$fully_participated_at == "")
```

```
## [1] 35142
```

From the 37,396 enrolments over the seven repetitions, there is 35,142 blank cells in the fully participated column. To remove the blank cells from the column a new table was created using the following code:

```
enrolments_fully_part = select(combined_enrolments, learner_id, rep, gender, age_range, fully_participa
  na_if("")%>%
  na.omit(combined_enrolments$fully_participated_at)
```

```
glimpse(enrolments_fully_part)
```

```
## Rows: 2,151
## Columns: 9
## $ learner_id               <chr> "ecdd37db-0c75-496e-bff2-230553d0e38c", "25cc3~
## $ rep                      <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
## $ gender                   <chr> "Unknown", "Unknown", "Unknown", "Unknown", "m~
## $ age_range                <chr> "Unknown", "Unknown", "Unknown", "Unknown", "3~
## $ fully_participated_at    <chr> "2016-09-22 16:56:03 UTC", "2016-10-25 12:44:1~
## $ highest_education_level <chr> "Unknown", "Unknown", "Unknown", "Unknown", "s~
## $ employment_status        <chr> "Unknown", "Unknown", "Unknown", "Unknown", "w~
## $ employment_area          <chr> "Unknown", "Unknown", "Unknown", "Unknown", "a~
## $ detected_country         <chr> "NG", "NO", "GB", "GB", "IT", "UA", "KE", "GB"~
```

This leaves 2151 learners that have fully completed the course. It was then possible to create a new table where the learner id of the 2151 students was used to link this table with the question responses of these learners. This table can be seen below:

```
glimpse(fully_participated_learners_qrespones)
```

```
## Rows: 71,132
## Columns: 19
## $ learner_id               <chr> "ecdd37db-0c75-496e-bff2-230553d0e38c", "ecdd3~
## $ rep.x                    <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
## $ gender                   <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ age_range                <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ fully_participated_at    <chr> "2016-09-22 16:56:03 UTC", "2016-09-22 16:56:0~
## $ highest_education_level <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ employment_status        <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ employment_area          <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ detected_country         <chr> "NG", "NG", "NG", "NG", "NG", "NG", "NG", "NG"~
## $ quiz_question            <chr> "1.7.1", "1.7.2", "1.7.2", "1.7.3", "1.7.4", "~
## $ question_type            <chr> "MultipleChoice", "MultipleChoice", "MultipleC~
## $ week_number              <int> 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 2, 2, 2, 2~
## $ step_number              <int> 7, 7, 7, 7, 7, 7, 7, 7, 11, 11, 11, 11, 8, 8, ~
## $ question_number          <int> 1, 2, 2, 3, 4, 5, 6, 6, 1, 1, 2, 3, 1, 1, 1, 2~
## $ response                 <chr> "1,2,3", "1", "2", "1,2,3,4,5", "2", "2", "2",~
## $ cloze_response           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ submitted_at             <chr> "2016-09-13 01:23:52 UTC", "2016-09-13 01:25:1~
## $ correct                  <chr> "true", "false", "true", "true", "true", "true~
## $ rep.y                    <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
```

There are still columns in this table which aren't helpful when focusing on the the students responses to questions. As demonstrated earlier in this report the "cloze_response" column contains NA in every row so therefore can clearly be removed. The submitted at column also holds no real relevance to how a student responds to a question. Also we can see in the merging of the the two tables the repetition or rep column has been duplicated as it was in both tables, so this can also be removed. The following code will be added to remove the columns

```
fully_participated_learners_qrespones = left_join(enrolments_fully_part, qresponse_all, by = "learner_i
  fully_participated_learners_qrespones = select(fully_participated_learners_qrespones, -cloze_response
  group_by(learner_id)
```

```
glimpse(fully_participated_learners_qrespones_clean)
```

```
## Rows: 71,132
## Columns: 16
## Groups: learner_id [2,145]
## $ learner_id              <chr> "ecdd37db-0c75-496e-bff2-230553d0e38c", "ecdd3~
## $ rep.x                   <chr> "rep 1", "rep 1", "rep 1", "rep 1", "rep 1", "~
## $ gender                  <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ age_range               <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ fully_participated_at   <chr> "2016-09-22 16:56:03 UTC", "2016-09-22 16:56:0~
## $ highest_education_level <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ employment_status       <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ employment_area         <chr> "Unknown", "Unknown", "Unknown", "Unknown", "U~
## $ detected_country        <chr> "NG", "NG", "NG", "NG", "NG", "NG", "NG", "NG"~
## $ quiz_question           <chr> "1.7.1", "1.7.2", "1.7.2", "1.7.3", "1.7.4", "~
## $ question_type           <chr> "MultipleChoice", "MultipleChoice", "MultipleC~
## $ week_number             <int> 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 2, 2, 2, 2~
## $ step_number             <int> 7, 7, 7, 7, 7, 7, 7, 7, 11, 11, 11, 11, 8, 8, ~
## $ question_number         <int> 1, 2, 2, 3, 4, 5, 6, 6, 1, 1, 2, 3, 1, 1, 1, 2~
## $ response                <chr> "1,2,3", "1", "2", "1,2,3,4,5", "2", "2", "2",~
## $ correct                 <chr> "true", "false", "true", "true", "true", "true~
```

Format data:

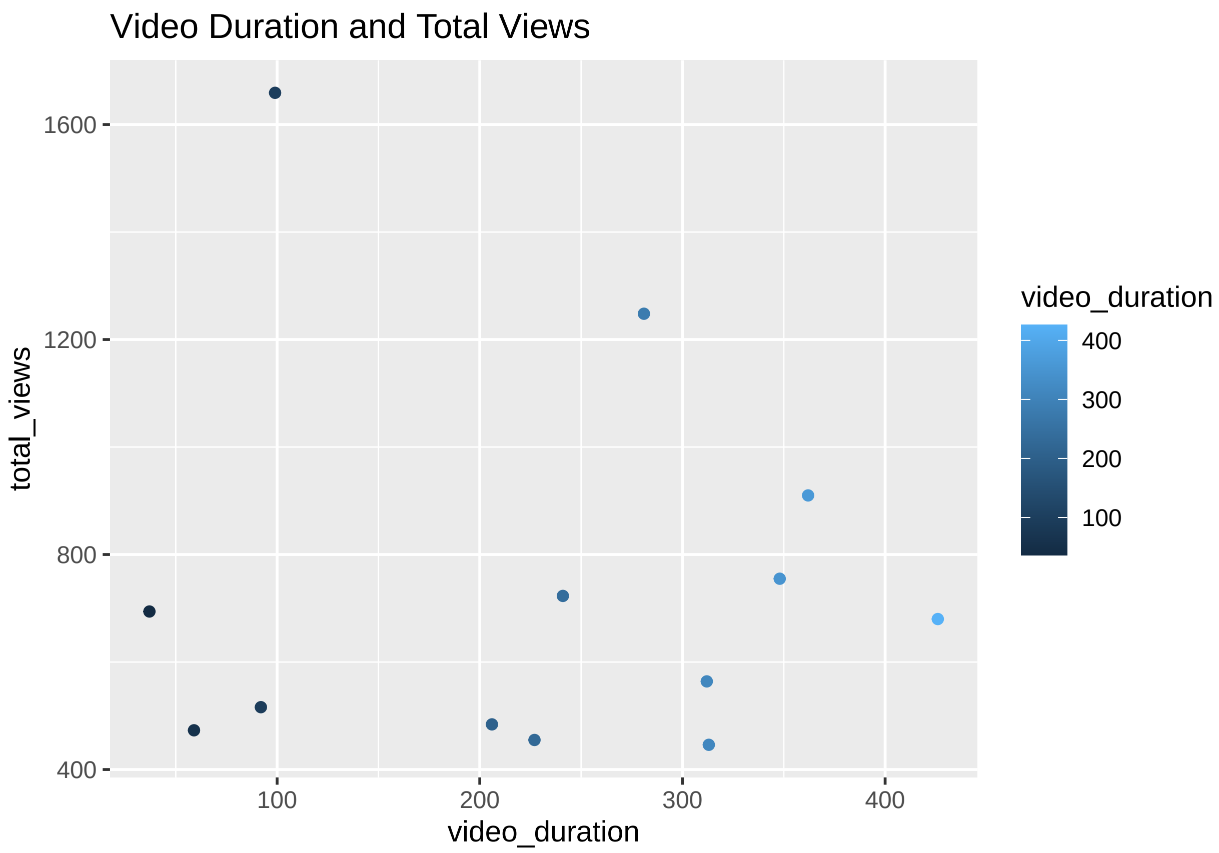
Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.

IV. Modeling

What is widely regarded as data science's most exciting work is also often the shortest phase of the project.

Here you'll likely build and assess various models based on several different modeling techniques. This phase has four tasks:

```
ggplot(data = videostats_yr1, mapping = aes(x = video_duration, y = total_views)) +
geom_point(aes(colour = video_duration)) +
ggtitle("Video Duration and Total Views")
```



Select modeling techniques:

Determine which algorithms to try (e.g. regression, neural net).

Generate test design:

Pending your modeling approach, you might need to split the data into training, test, and validation sets.

Build model:

As glamorous as this might sound, this might just be executing a few lines of code like "reg = LinearRegression().fit(X, y)".

Assess model:

Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.

Although the CRISP-DM guide suggests to "iterate model building and assessment until you strongly believe that you have found the best model(s)", in practice teams should continue iterating until they find a "good enough" model, proceed through the CRISP-DM lifecycle, then further improve the model in future iterations.

V. Evaluation

Whereas the Assess Model task of the Modeling phase focuses on technical model assessment, the Evaluation phase looks more broadly at which model best meets the business and what to do next. This phase has three tasks:

Evaluate results:

Do the models meet the business success criteria? Which one(s) should we approve for the business? ### Review process: Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.

Determine next steps:

Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

VI. Deployment

“Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.”

–CRISP-DM Guide

A model is not particularly useful unless the customer can access its results. The complexity of this phase varies widely. This final phase has four tasks:

Plan deployment:

Develop and document a plan for deploying the model.

Plan monitoring and maintenance:

Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model.

Produce final report:

The project team documents a summary of the project which might include a final presentation of data mining results.

Review project:

Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.