CSC8631: Data Management and Exploratory Data Analysis

Simon Irvine | 210449787

03 December, 2021

Assignment

Data Management and Exploratory Data Analysis - CSC8631 Coursework (Semester 1, 2021)

- Module Leader: Dr Matthew Forshaw
- Lecturer: Dr Joe Matthews

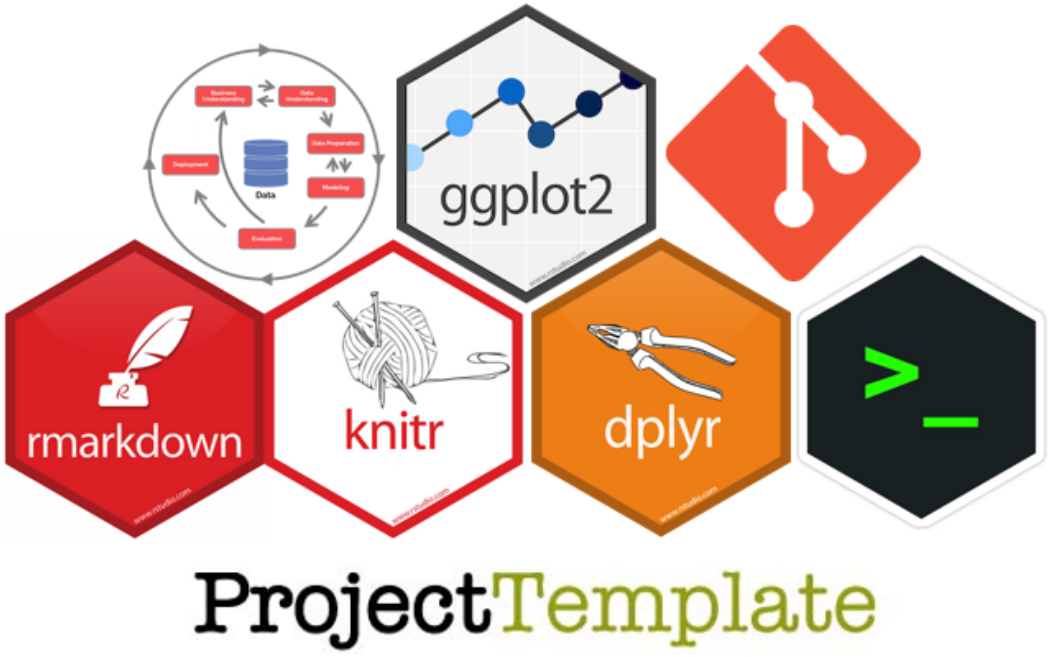


Figure 1: Data management and exploratory data analysis tools

The Brief

Scenario

Learning Analytics, a rapidly-growing application area for Data Science, is defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environment in which it occurs"[1].

Existing mechanisms to record student engagement (e.g. attendance monitoring) fail to capture the extent and quality of engagement outside of the classroom environment. Further complementary sources of data are routinely collected about our learners (e.g. use of on-campus facilities, Virtual Learning Environment (VLE) and Re-Cap access, and student wellbeing referrals); however, these currently reside in a number of silos.

Learning Analytics seeks to aggregate these sources of data to derive shared insights, and provide effective measures of engagement. Insights may inform learning design, inform intervention processes for at-risk students, and improve student attainment.

The most complete introduction is available in government policy policy report "From Bricks To Clicks"[2]. The report is quite extensive, but there are some nice case studies from Nottingham Trent and the OU to give you a flavour of the types of projects in this area.

Challenge

In this project we will emulate a very familiar process undertaken by data analysts. We will take a dataset provided to us, and develop a suite of tools which allow us to extract interesting insights from this data in a quick, reliable and repeatable manner. The datasets you are expected interpret as a data analyst are commonly previously unseen, so the process of building a pipeline is an exploratory one. Consequently, you will be expected to review and interrogate the data to gain an understanding of its structure and composition.

In this coursework you will develop a data analysis pipeline to explore a given dataset. There are no formal requirements for the functionality or focus of your analysis. Your data analysis should follow routes of enquiry which are of greatest interest to you. Therefore, there exists scope for a great deal of flexibility so we anticipate solutions to this challenge will vary.

We encourage you to pursue ambitious analysis, but just as importantly we are looking for good programming practice. When developing large systems such as these, it is important that you write your code incrementally, and test it carefully before continuing to add additional functionality.

[1] George Siemens and Phil Long. Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE review, 46(5):30, 2011

[2] http://www.policyconnect.org.uk/hec/sites/site_hec/files/report/419/fieldreportdownload/frombrickstoclicks-hecreportforweb.pdf

Business Understanding

In addition to providing an exploratory data analysis of a given dataset, best-practice development is of key relevance to successfully completing the assignment to a sufficient quality standard. Therefore, as we explore the dataset - asking questions and iteratively diving deeper - the author will highlight goals, objectives and success criteria that will relate to the actual analysis work **and** to version control, documentation, reproducibility, and 'literate programming'.

Objectives

The primary objective is to develop a data analysis pipeline to explore the provided dataset maintaining a balance between being an *interative* and *creative* process Since there is less emphasis on the generating successful findings, the author considers an interative scientific method to be most appropriate.

Secondarily, an important objective is to present this analysis in a well-documented and reproducible manner - allowing other analysts to recreate and further develop this work with minimal frustration or with diverging results.

Success Criteria

Based on these main objectives, the author proposes the following simple success critera for this exploratory data analysis:

- Data are imported and processed allowing exploratory data analysis;
- A repository and version control system is established;
- Appropriate questions are asked throughout to further analysis;
- Processes and code are understandable and reproducible;
- A final report is produced.

Situation

For this analysis it is important to consider the resources available: personnel, data, computing resources, and software.

Inventory of Resources

1. **Personnel**: Author, peer study group, and academic lecture team
2. **Data**: Dataset MOOC FutureLearn Cybersecurity - admin logs (62 .csv files) and course overview documents (7 .pdf files)
3. **Computing**: Personal laptop (AMD Ryzen 7 with 16GB RAM), and Newcastle University Tier 2 Azure Virtual Desktop
4. **Software**: R Studio (R, projecttemplate, dplyr, ggplot, readr), Git Bash, PowerPoint

Assumptions and Constraints

The author's initial assumptions falls into two main categories; a) technical, and b) practical. The technical assumptions are as follows:

- The instruction and guidance provided by the academic staff is sufficient to be able to understand and *solve* the problem of this assignment; and,
- The proposed tools and processes, namely CLI, R (including packages) and CRISP-DM, are sufficient to successfully complete this assignment.

The practical assumptions are as follows:

- The dataset quality is sufficient without support from an subject matter expert to navigate any noise or bias;
- There are insights to be gained from the dataset worth of this exploratory analysis; and,
- The author possesses the skill to leverage the tools appropriately to allow successful analysis.

The constraints for this piece of work are largely related to the author's skillset (with no prior experience of using these tools) and sufficient time allocation to the analysis. Although preferable, it is not necessary for this analysis to provide any actionable insights or successful findings - however, its processes, documentation and reprodicibility are important.

Risks and contingencies

There are no specific risks to this analysis, other than that there is a submission deadline. Illness, poor time management or an act of God can be mitigated by careful planning and updating the customer (academic staff) of an increased likelihood of missing the deadline.

Terminology

It is useful to provide a brief glossary of terms used within data analytics projects to support business understanding.

Table 1: Glossary of Business Terms

Term	Definition
Cybersecurity	The protection of computer systems and networks from information disclosure, theft of or damage to their hardware, software, or electronic data, as well as from the disruption or misdirection of the services they provide.
FutureLearn	A British digital education platform founded in December 2012. The company is jointly owned by The Open University and SEEK Ltd. It is a Massive Open Online Course, ExpertTrack, microcredential and Degree learning platform.
Learner ID	A variable used within the dataset to identify a user while maintaining their anonymity.

Term	Definition
MOOC	A massive open online course is an online course aimed at unlimited participation and open access via the Web.
Sentiment survey	A weekly questionnaire completed by users to gain feedback on the user and learning experience.

Table 2: Glossary of Data Analytics Terms

Term	Definition
Algorithm	An unambiguous mathematical specification or statistical process used to perform analysis of data.
Correlation	Measure of association of two variables.
Dataset	A dataset is the base of all multivariate data analysis, often also called a data matrix. It is made up of values of several different variables for a number of observations.
Data analytics	The process of examining large data sets to uncover hidden patterns, unknown correlations, trends, customer preferences and other useful business insights.
Data science	A discipline that combines statistics, data visualization, computer programing, data mining and software engineering to extract knowledge and insights from large and complex data sets.
dplyr	An open-source data manipulation package for the statistical programming language R.
ggplot	An open-source data visualization package for the statistical programming language R.
Histogram	A column (bar) plot visualizing the distribution of a variable.
Linear regression	A statistical method used to summarize and show relationships between variables.
Outliers	Extreme values that might be errors in measurement and recording, or might be accurate reports of rare events.
R	A programming language and free software environment for statistical computing and graphics. It is widely used among statisticians and data miners for developing statistical software and data analysis.
Version control	The practice of tracking and managing changes to software code. Version control systems are software tools that help software teams manage changes to source code over time.

Data Mining Goals

Since *success findings* are not a key requirement of the customer, the data mining goals do not require customary specificity. However, it is the goal of this analysis to gain at least (3) insights from the dataset relating to user experience and completion of the online course **'Cyber Security: Safety at Home, Online, in Life'** during its multiple deliveries (runs) over a two-year period.

Project Plan

For a exploratory data analysis project such as this, with a minimal number of resources, only a simple plan is required to achieve the success criteria.

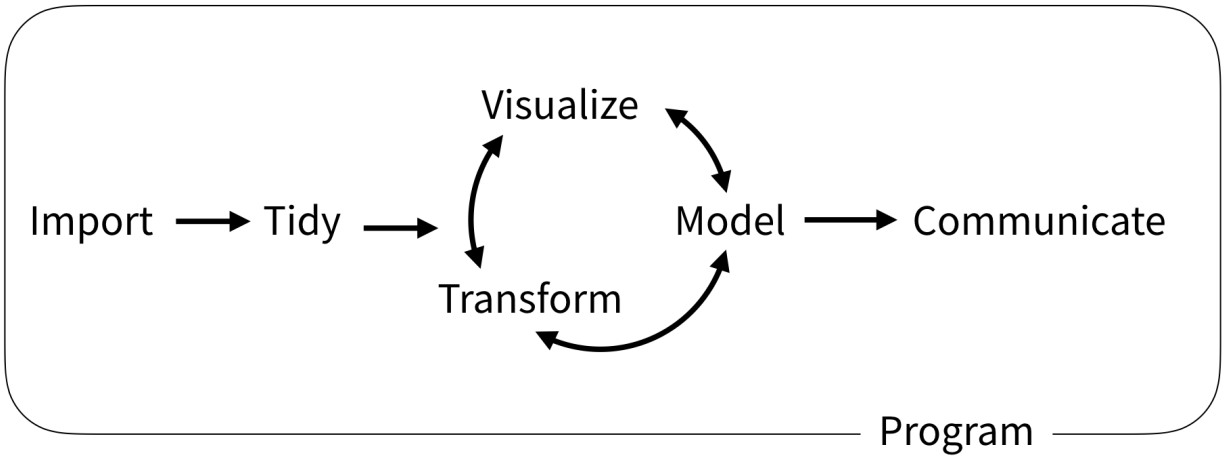


Figure 2: The **"R for Data Science"** model of required tools[3]

It is the intention of the author to employ a plan based on CRISP-DM[4] to set *guardrails* to help plan, organize and implement this project. A relevant distillation of this can be found in the "R for Data Science" model (see Figure 2) with the steps included within the interation cycle. The author will use this model as the basis for his plan, as follows:

1. **Import**

- Preparing project structure
- Setting up environment and packages
- Collecting data

2. **Tidy**

- Exploring data
- Pre-processing and cleaning

3. **The Loop**

- Questioning
- Transforming
- Visualizing
- Insights

4. **Communicate**

- Report
- Presentation
- Future work

[3]R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Whickham, H and Grolemund, G. O'Reilly Media; 1st edition (January 17, 2017)

[4]**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) is a process model with six phases that naturally describes the data science life cycle.

1. Import

For this analysis the dataset has already been provided with no specific requirements to perform technical data collection processes. The import process will consist mostly of establishing an opinionated project structure, setting up the analysis environment and adding the dataset to the working directory.

Preparing project structure

As per best practice, the author has elected to use a semi-automated project tool that will create the working directory folders, config files and run pre-processing scripts. This package is compatible with the chosen IDE (*R Studio*) and is called *ProjectTemplate*[5].

Setting up environment and packages

Since the chosen language is R, the author has chosen to use *R Studio*[6] and install the *TidyVerse*[7] package, which comprises multiple data manipulation and graphics libraries, suitable for this kind of analysis.

Environment and Packages

Initially, we need to install the required R packages to allow us to perform our analysis. We are most interested manipulating data, tidying it up and creating graphics.

```
# Install R packages
install.packages("tidyverse")
install.packages("projecttemplate")
```

Project Template

Creating a project with ProjectTemplate automatically populates a structured directory for us to work within.

```
# Let's set up the project using Project Template
  library("ProjectTemplate")
  create.project("CSC8631-SIrvine")
```

Git

Version control is a system that records changes to a file or set of files over time so that you can recall specific versions later. This is important for the integrity and reproducibility of a project. The author has selected Git to provide version control for this analysis project - with commits also pushed to Github.

[5]ProjectTemplate is a system for automating the thoughtless parts of a data analysis project. See http://projecttemplate.net/getting_started.html for details on how to set up.

[6]See https://www.rstudio.com/

[7]TidyVerse is an opinionated collection of R packages designed for data science. See https://www.tidyverse.org/packages/ for more information.

A Git directory is where Git stores the metadata and object database for your project. This is the most important part of Git, and it is what is copied when you clone a repository from another computer.

The basic Git workflow goes something like this:

1. You modify files in your working tree.
2. You selectively stage just those changes you want to be part of your next commit, which adds only those changes to the staging area.
3. You do a commit, which takes the files as they are in the staging area and stores that snapshot permanently to your Git directory.

Within a Command Line Interface, i.e. Git Bash:

```
# Change directory to where your project is based, i.e. the working directory
  $ cd /to/path/

# Lists all files and folders in the directory
  $ ls

# Adds all files in the directory to the git
  $ git add -A

# This is where you save each version of your work
# You can add comments in case you need to roll back later
  $ git commit -m "Add comments here"

# Optionally you can push your local Git to a server, such as Github.
# Set up repository and push commits (versions) later.
  $ git push -u original main
```

Collecting data

The author received the dataset and copied the the zip file over to the *data folder* of the working directory.

Within a Command Line Interface, i.e. Git Bash:

```
unzip -d /CSC8631-SIrvine/data/ {FutureLearn MOOC Dataset.zip}
```

2. Tidy

Tidying refers to a systematic and opinionated method of organizing data to support data science tasks. Since we are using *TidyVerse*, we will utilize the *tidyr* package to tidy data. Once you have tidy data and the tidy tools provided by packages in the tidyverse, you will spend much less time munging data from one representation to another, allowing you to spend more time on the analytic questions at hand.

Exploring data

As an initial step we will explore the data using basic tools within the R package.

```
# Define data folder location and target files. This will prompt you to target the data
↪  folder.
data_folder = choose.dir()
files = list.files(path = data_folder)

# Number of files
length(files)
```

```
## [1] 62
```

```
# Top 15 files in data folder
head(files, 15)
```

```
##  [1] "5 step names overview.html"                                "cyber-security-1_archetype-survey-resp
##  [3] "cyber-security-1_enrolments.csv"                           "cyber-security-1_leaving-survey-respon
##  [5] "cyber-security-1_question-response.csv"                    "cyber-security-1_step-activity.csv"
##  [7] "cyber-security-1_weekly-sentiment-survey-responses.csv"    "cyber-security-2_archetype-survey-resp
##  [9] "cyber-security-2_enrolments.csv"                           "cyber-security-2_leaving-survey-respon
## [11] "cyber-security-2_question-response.csv"                    "cyber-security-2_step-activity.csv"
## [13] "cyber-security-2_team-members.csv"                         "cyber-security-2_weekly-sentiment-surv
## [15] "cyber-security-3_archetype-survey-responses.csv"
```

```
# Let's have a look at one of these .csv files
library(dplyr)
head(cyber.security.1_enrolments, 10)
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 10 x 13
##    learner_id enrolled_at unenrolled_at role  fully_participa~ purchased_state~ gender country age_ra
##    <chr>      <chr>       <chr>         <chr> <chr>           <chr>            <chr>  <chr>   <chr>
##  1 160d6600-~ 2016-08-10~ ""            lear~ ""              ""               Unkno~ Unknown Unknow
##  2 4dc22fed-~ 2016-05-24~ "2018-10-30 ~ lear~ ""              ""               male   PE      46-55
```

```
##  3 ecdd37db-~ 2016-05-19~ ""            lear~ "2016-09-22 16:~ ""              Unkno~ Unknown Unknou
##  4 988964c9-~ 2016-05-19~ ""            lear~ ""               ""              Unkno~ Unknown Unknou
##  5 f1493366-~ 2016-09-19~ ""            lear~ ""               ""              Unkno~ Unknown Unknou
##  6 25cc3b46-~ 2016-08-30~ ""            lear~ "2016-10-25 12:~ ""              Unkno~ Unknown Unknou
##  7 9c23a086-~ 2016-06-22~ ""            lear~ "2016-10-10 11:~ ""              Unkno~ Unknown Unknou
##  8 8851dc49-~ 2016-08-07~ ""            lear~ "2018-10-17 18:~ ""              Unkno~ Unknown Unknou
##  9 a59b0a12-~ 2016-08-02~ "2018-10-17 ~ lear~ ""               ""              Unkno~ Unknown Unknou
## 10 198c1017-~ 2016-09-09~ ""            lear~ ""               ""              Unkno~ Unknown Unknou
## # ... with 1 more variable: detected_country <chr>
```

As can be seen, there are over 60 files comprising *.csv* and *.pdf* formats. Based on a quick check of the files, we can see that these relate to a online MOOC, called **'Cyber Security: Safety at Home, Online, in Life'**[8] delivered by **FutureLearn** over a period of two years and seven runs.

At this stage, we might start asking what does the data in these dataframes refer to. Are there any interesting column lables (variables) within the dataset? From a quick exploration we can say a few things about the data:

- There were seven recorded runs of this MOOC;
- There are learners from many countries;
- Not all the learners completed the leaving or weekly sentiment surveys;
- A lot of the learner details are unspecified/unknown;
- Some videos were more popular than others.

Pre-processing and cleaning

As part of *ProjectTemplate*, we have some pre-processing already established such as loading appropriate libraries, calling cached objects and running munge scripts.

```
# Pre-processing script for MOOC FutureLearn Dataset EDA

# Libraries import
    library(tidyverse)
    library(ProjectTemplate)
```

Data cleaning is an important step to raise the data quality to level required for analysis. This may involve selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modeling. As a first step, the author proposes to join the runs into single dataframes using *rbind()*. An example is shown below.

```
# The dataset comprises .csv  and .pdf files related to the delivery (7 runs) of a MOOC by FutureLearn.
# For preprocessing, we want combine all the dataframes to allow easier analysis.

all.archetype.survey.responses =
  rbind.data.frame(
  cyber.security.1_archetype.survey.responses,
  cyber.security.2_archetype.survey.responses,
```

[8] This MOOC is still available, as of December 2021, on FutureLearn's website and offered as a partnership with Newcastle University. https://www.futurelearn.com/courses/cyber-security

```
  cyber.security.3_archetype.survey.responses,
  cyber.security.4_archetype.survey.responses,
  cyber.security.5_archetype.survey.responses,
  cyber.security.6_archetype.survey.responses,
  cyber.security.7_archetype.survey.responses)
```

Then we want to remove clearly irrelevant or incorrect columns, possibly modifying columns names if appropriate.

```
# Remove unwanted columns from dataframes

# Remove "purchased_statement_at"
all.enrolments = select(all.enrolments, learner_id, enrolled_at,    unenrolled_at,  role,
↪  fully_participated_at, gender,  country,    age_range,  highest_education_level,
↪  employment_status,  employment_area,    detected_country)
all.enrolments
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 34,954 x 12
##    learner_id    enrolled_at   unenrolled_at   role  fully_participa~ gender country age_range highes
##    <chr>         <chr>         <chr>           <chr> <chr>            <chr>  <chr>   <chr>     <chr>
##  1 160d6600-ea0~ 2016-08-10 1~ ""              lear~ ""               Unkno~ Unknown Unknown   Unknou
##  2 4dc22fed-63d~ 2016-05-24 1~ "2018-10-30 20~ lear~ ""               male   PE      46-55     unive
##  3 ecdd37db-0c7~ 2016-05-19 0~ ""              lear~ "2016-09-22 16:~ Unkno~ Unknown Unknown   Unknou
##  4 988964c9-741~ 2016-05-19 2~ ""              lear~ ""               Unkno~ Unknown Unknown   Unknou
##  5 f1493366-17a~ 2016-09-19 1~ ""              lear~ ""               Unkno~ Unknown Unknown   Unknou
##  6 25cc3b46-a95~ 2016-08-30 0~ ""              lear~ "2016-10-25 12:~ Unkno~ Unknown Unknown   Unknou
##  7 9c23a086-f6b~ 2016-06-22 1~ ""              lear~ "2016-10-10 11:~ Unkno~ Unknown Unknown   Unknou
##  8 8851dc49-028~ 2016-08-07 1~ ""              lear~ "2018-10-17 18:~ Unkno~ Unknown Unknown   Unknou
##  9 a59b0a12-af4~ 2016-08-02 1~ "2018-10-17 21~ lear~ ""               Unkno~ Unknown Unknown   Unknou
## 10 198c1017-51f~ 2016-09-09 2~ ""              lear~ ""               Unkno~ Unknown Unknown   Unknou
## # ... with 34,944 more rows
```

```
# Remove "id" and "responded_at"
all.archetype.survey.responses = select(all.archetype.survey.responses, learner_id,
↪  archetype)

# Remove "id"
all.archetype.survey.responses = select(all.leaving.survey.responses, learner_id,
↪  left_at,  leaving_reason, last_completed_step_at, last_completed_step,
↪  last_completed_week_number, last_completed_step_number)

# Remove "cloze_response", "question_type", "week_number",  "step_number",
↪  "question_number",  "response", "submitted_at"
all.question.response = select(all.question.response, learner_id,   quiz_question,
↪  correct)
```

```
# Remove "week_number", "step_number", "first_visited_at", "last_completed_at"
all.step.activity = select(all.step.activity, learner_id,   step)

# Remove "first_name", "last_name", "user_role" and change "id" to "learner_id" as this
↪  is an error
all.team.members = select(all.team.members, id, team_role)
all.team.members = rename(all.team.members, learner_id = id)

# Remove "total_downloads", "total_caption_views", "    total_transcript_views",
↪  "viewed_hd", "viewed_five_percent", "viewed_ten_percent",
↪  "viewed_twentyfive_percent", "viewed_seventyfive_percent",
↪  "viewed_onehundred_percent", "console_device_percentage" ,
↪  "desktop_device_percentage", "mobile_device_percentage", "tv_device_percentage",
↪  "tablet_device_percentage", "unknown_device_percentage",
↪  "antarctica_views_percentage"
all.video.stats = select(all.video.stats, step_position,    title,  video_duration,
↪  total_views,    viewed_fifty_percent,   viewed_ninetyfive_percent,
↪  europe_views_percentage,    oceania_views_percentage,   asia_views_percentage,
↪  north_america_views_percentage, south_america_views_percentage,
↪  africa_views_percentage)

# Remove "id" and "responded_at"
all.weekly.sentiment.response.surveys = select(all.weekly.sentiment.response.surveys,
↪  week_number,  experience_rating,  reason)
```

It may be necessary to revisit this cleaning and merging stage if we run into difficulties during the transforming stage.

3. The Loop

We now move onto the Transform<->Visualize<->Model Loop (aka *The Loop*) and will start by asking some questions that we can start to investigate.

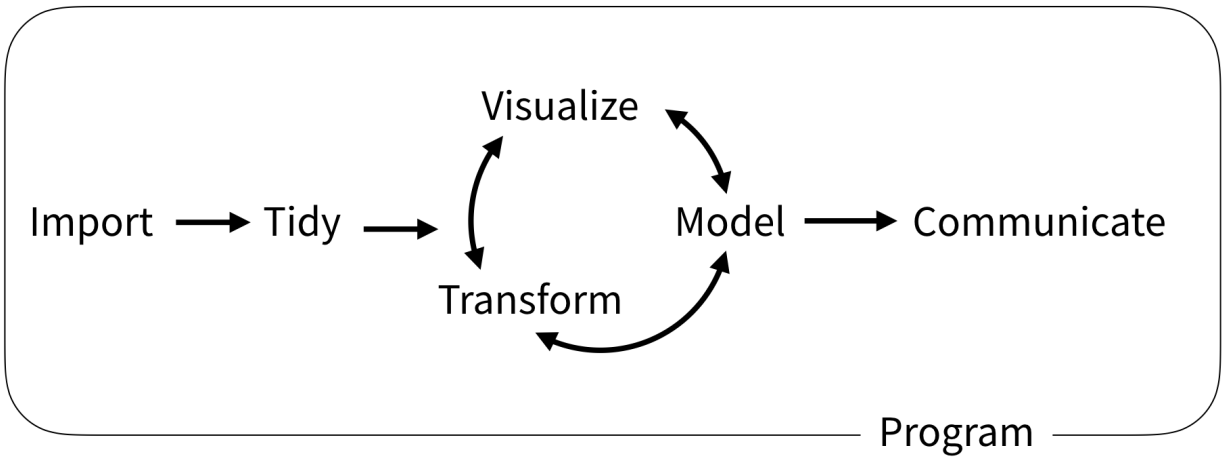


Figure 3: The **"R for Data Science"** model of required tools

Questionning

After exploring the data during the previous stages, we can ask some question worthy of investigation. These can be revised, changed or replaced during the process and it is expected to go through iterations (hence *The Loop*).

Q1. What can we learn about the people who registered for this MOOC?

We will need to look at dataframes that include information related to the users,if they completed the MOOC and reasons for leaving.

Transforming

Within the dataset, there is an opportunity to merge columns from different dataframes to aid exploratory analysis.

```
# Creating user_profiles dataframe to understand users' backgrounds, when they left the
↪  MOOC and why, and their learner archetype
user_profiles = left_join(all.enrolments,all.archetype.survey.responses, by = NULL, copy
↪  = FALSE, keep = FALSE)
```

```
## Joining, by = "learner_id"
```

```
user_profiles = left_join(user_profiles,all.leaving.survey.responses,by = NULL, copy =
↪  FALSE, keep = FALSE)
```

```
## Joining, by = c("learner_id", "left_at", "leaving_reason", "last_completed_step_at", "last_completed
```

```
# We will identify any FutureLearn team members within the user list.
user_profiles = left_join(user_profiles,all.team.members, by = NULL, copy = FALSE, keep =
↪  FALSE)
```

```
## Joining, by = "learner_id"
```

```
# Several values are displayed as *unknown* and this may cause issues when creating plots
↪  so we will replace them with *NA*
user_profiles = na_if(user_profiles,"Unknown")
user_profiles = as_tibble(user_profiles)
```

Visualizing

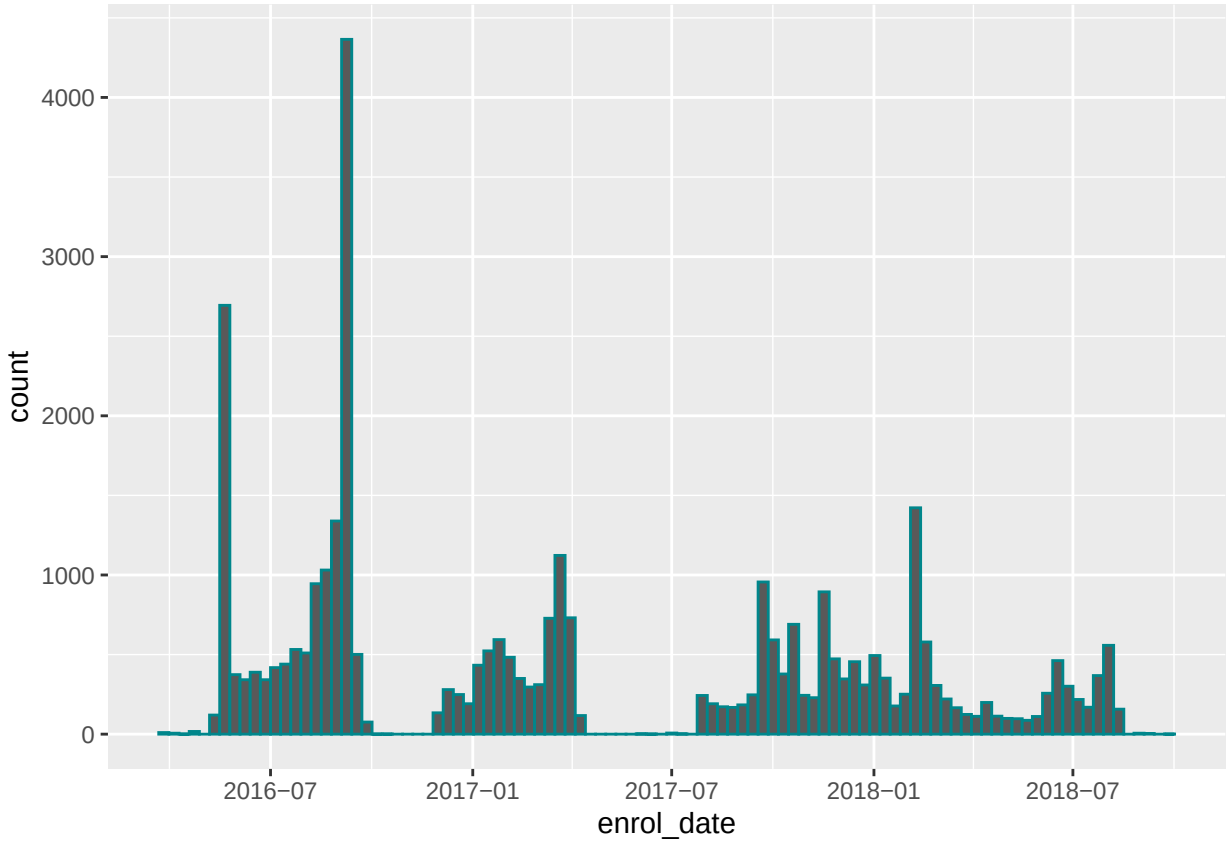
The first visualization we can create is a summary of enrolments on the MOOC.

```
library(ggplot2)
library(dplyr)

# We need to convert the dates into date class
enrol_date = as.Date(user_profiles$enrolled_at)
participate_date = as.Date(as_datetime(user_profiles$fully_participated_at))
leave_date = as.Date(user_profiles$left_at)

# Calculation to understand how long learners stay on the course or how long to
↪  'participate' - this was not used for the following plots.
time_to_participate = participate_date - enrol_date
time_to_leave = leave_date - enrol_date

# Plotting enrolment date counts
ggplot(user_profiles, aes(enrol_date)) + geom_histogram(color = 'turquoise4', bins = 100,
↪  show.legend = TRUE)
```



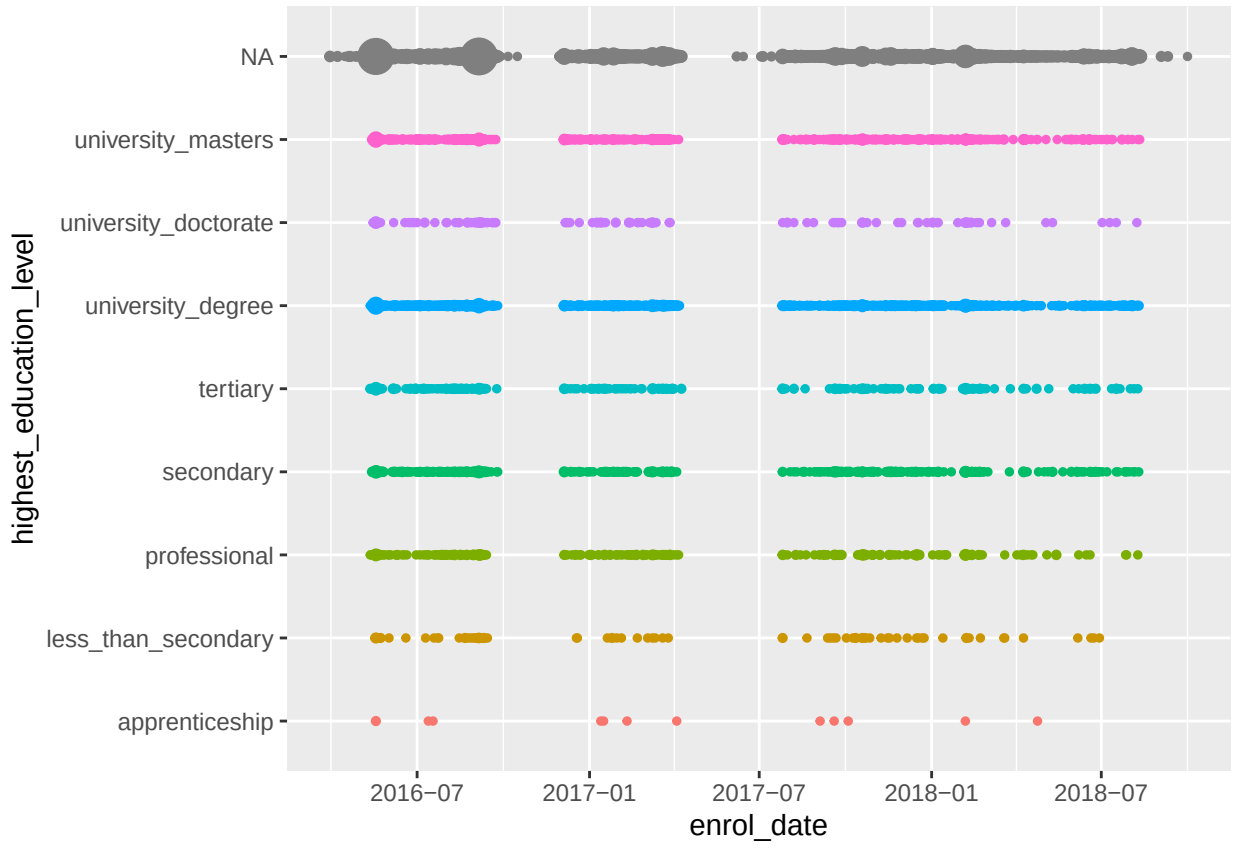
Insights

A basic visualization of enrolments against date. We can see that, although there are seven runs of the MOOC, there are distinct clusters of enrolments starting in 2016 and finishing up in 2018. Other questions that come to mind when looking at this barchart:

- Why are there such large spikes of enrolment in 2016? Was there a marketing drive?
- It would be good to know when the MOOC runs began and finished to delineate the groupings.
- It is clear registration was open after each course started.

```
library(ggplot2)
library(dplyr)

# Plotting enrolments against education level
ggplot(user_profiles, aes(enrol_date,highest_education_level)) + geom_count(aes(colour =
↪  factor(highest_education_level)),show.legend = FALSE)
```



Insights

This chart gives an overview of the enrolments over time against highest education level (including NA).

- What does NA mean in this context? No education (unlikely, as secondary is an option) or learners did not complete the form.
- Overall it clear there were fewer outliers (PhD holders and less than secondary) which is not surprising. There seems to be a fairly even distribution across all other education levels.

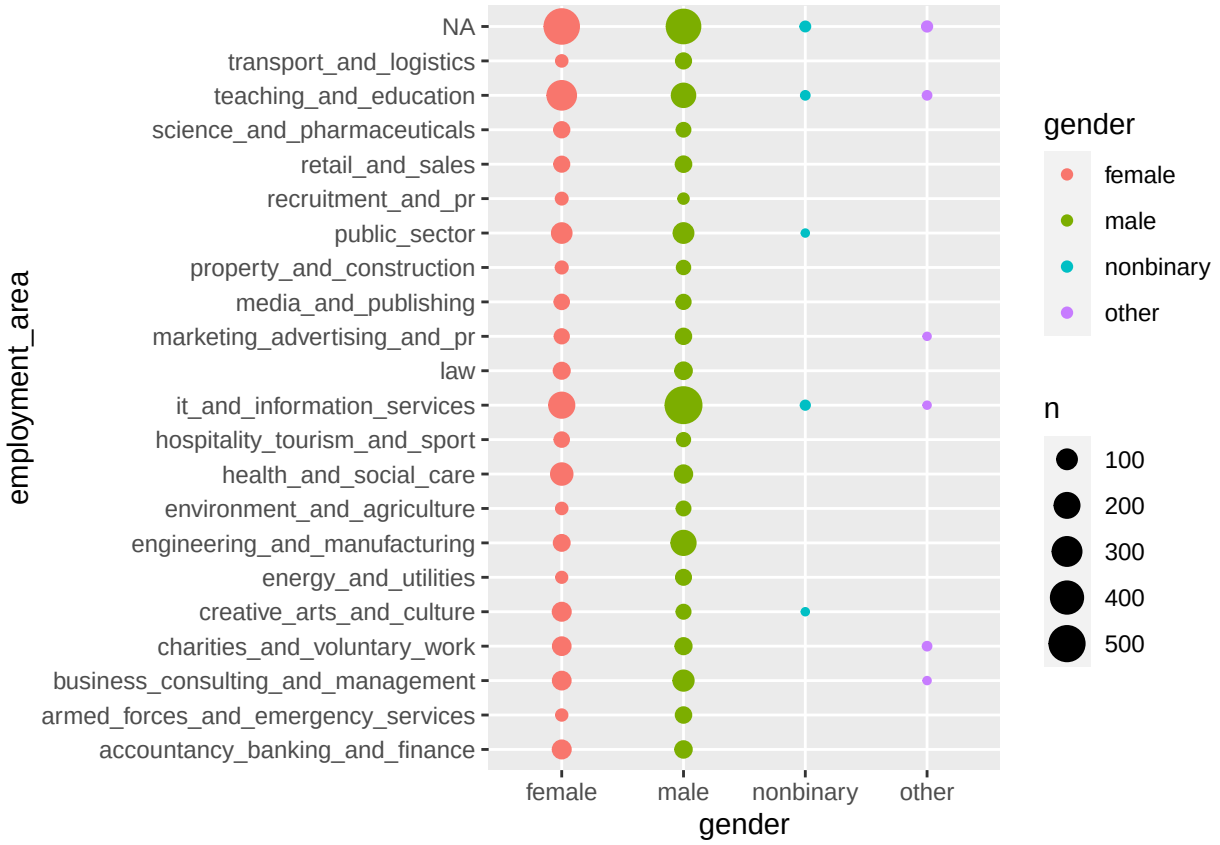
Q2. Does gender play a role?

A factor of interest is that of gender. Given the sample size and relative anonymity of the learners, could there be anything interesting to learn?

```
library(ggplot2)
library(dplyr)

# Clean gender column to remove NAs
user_profiles_complete = user_profiles[complete.cases(user_profiles$gender),]

# Plot of job types against gender
ggplot(user_profiles_complete, aes(gender, employment_area)) + geom_count(aes(colour =
↪  gender))
```



Insights

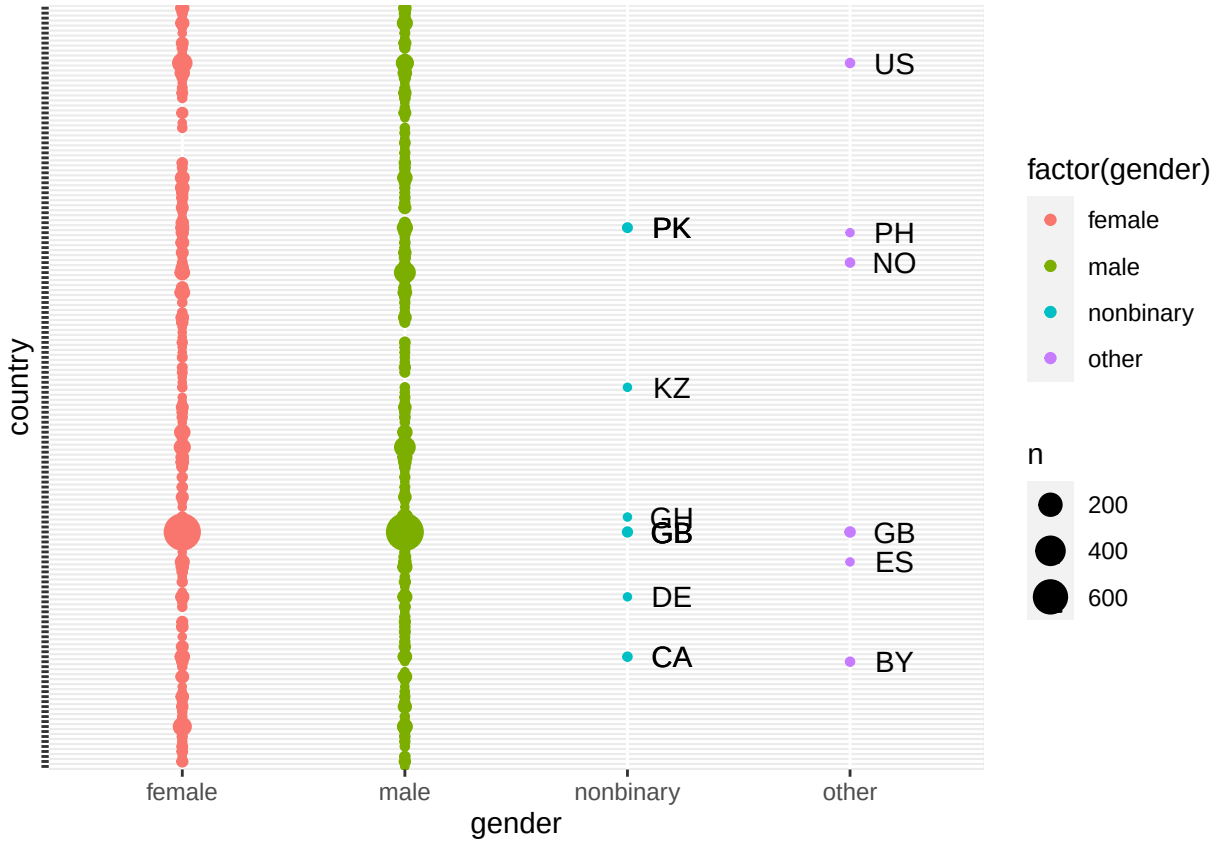
This chart presents employment area against gender with a NA category for those not in work or who preferred not to answer.

- There is a slightly greater number of females working in *teaching and education* and *accounting* on the course.
- There is a noticably greater number of males working in *IT and information services* and *engineering* on the course.
- Most other categories have an equal distribution.

- Nonbinary and other genders are represented in *teaching*, *marketing*, *IT*, *voluntary* and *business consulting*.

```
library(ggplot2)
library(dplyr)

# Plotting gender (count) against country
ggplot(user_profiles_complete, aes(gender,country)) + geom_count(aes(colour =
↪  factor(gender)),show.legend = TRUE) +theme(axis.text.y=element_blank()) +
↪  geom_text(data=subset(user_profiles_complete, gender == "nonbinary"), aes(label =
↪  country, size = NULL), nudge_x = 0.2) + geom_text(data=subset(user_profiles_complete,
↪  gender == "other"), aes(label = country, size = NULL), nudge_x = 0.2, check_overlap =
↪  TRUE)
```



Insights

This chart presents genders against country as learners on the MOOC. The author considered it interesting to see if there were any surprises in terms on nonbinary and other gender identification.

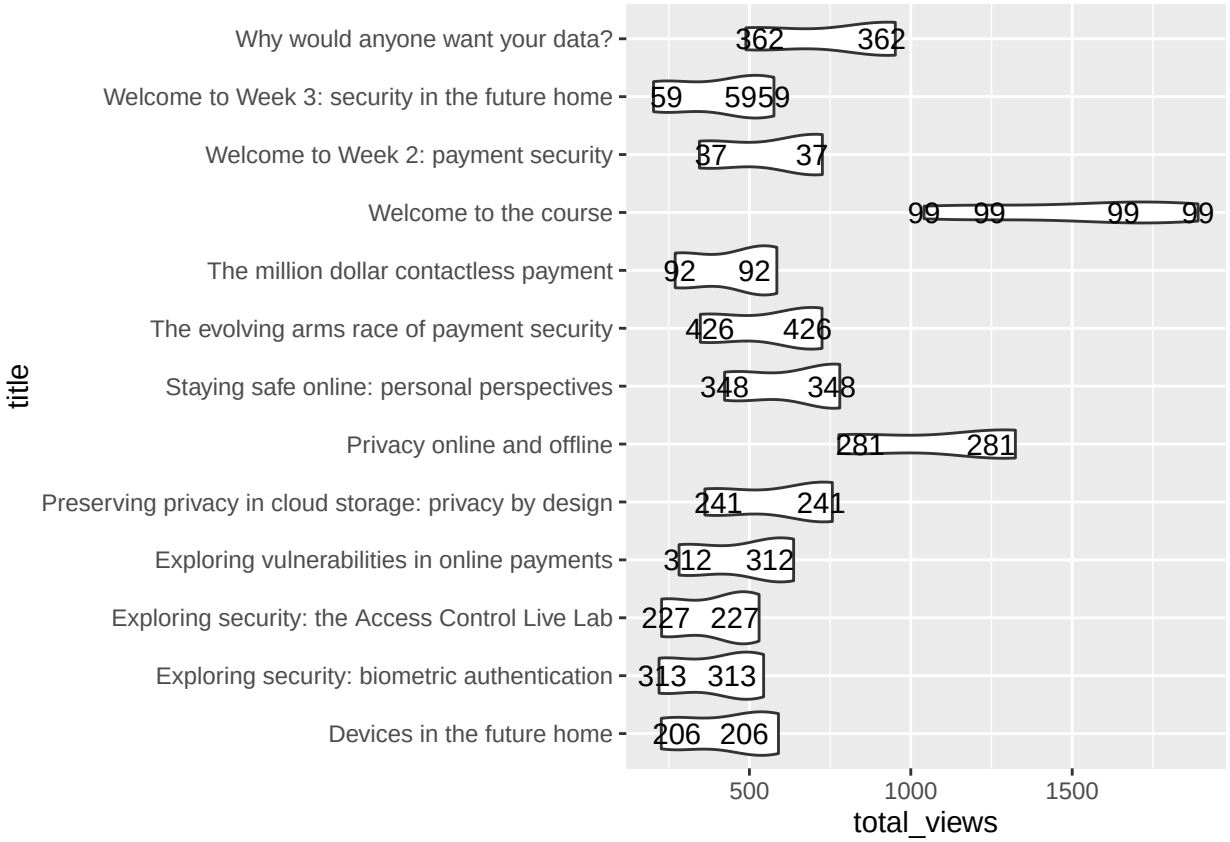
- Nonbinary learners registered home countries as Pakistan, Kazakhstan, Ghana, UK, Germany and Canada.
- Other gender learners registered home countries as US, Norway, Philippines, UK, Spain and Bulgaria.
- Unsurprisingly, regardless of gender, UK learners represented the largest group.

Q3. How popular were the videos?

A short exploration of the video content on the MOOC.

```
library(ggplot2)
library(dplyr)

# A violin chart of videos on the MOOC dispaying the total views
ggplot(all.video.stats, aes(total_views,title)) + geom_violin()  + geom_text(aes(label =
↪  video_duration, size = NULL),check_overlap = TRUE, nudge_x = 0.2)
```



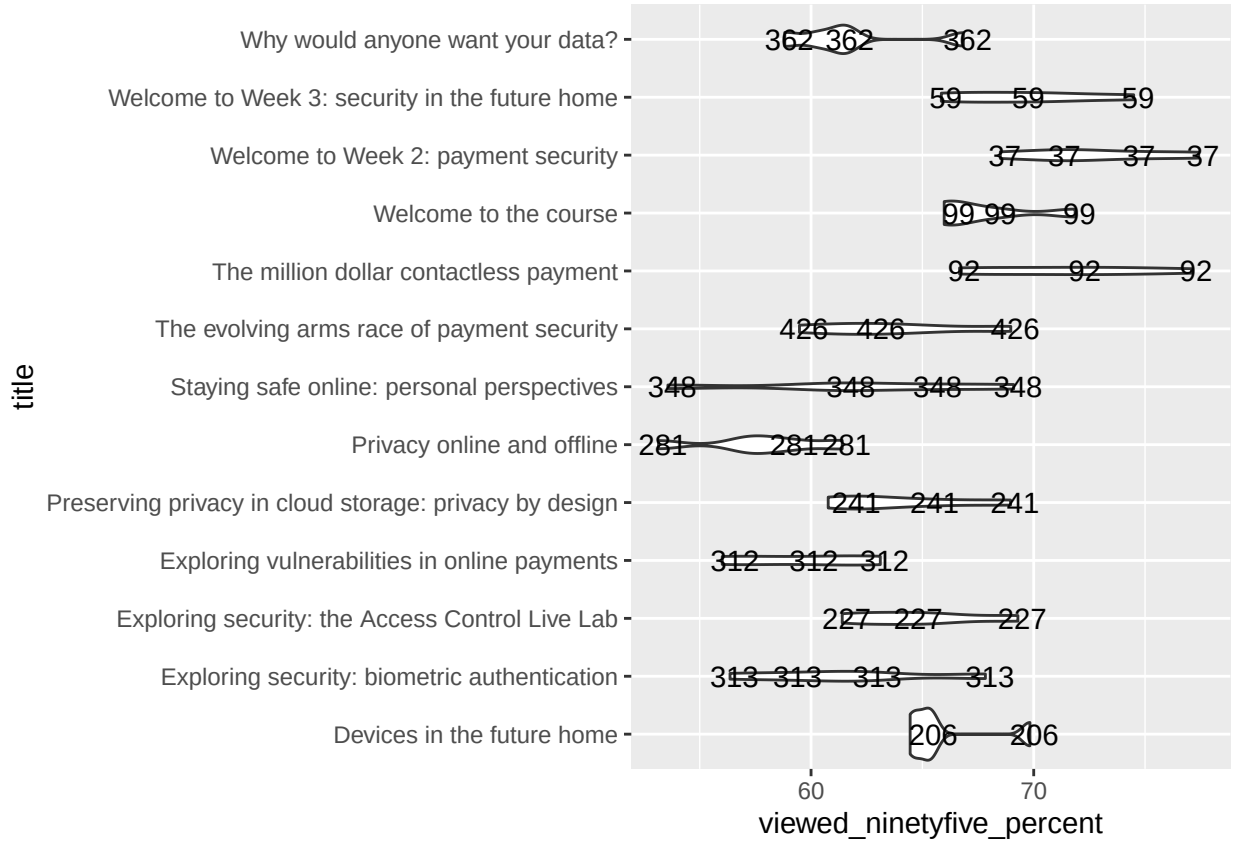
Insights

This chart presents total views across all videos on the MOOC.

- The *Welcome to the course* video is the most watched video followed by *Privacy online and offline*.
- There is little variation across the other videos with a 2% watch rate.

```
library(ggplot2)
library(dplyr)

# Let us look at what video engagement was like - how many people watched at least 95% of
↪  the video?
ggplot(all.video.stats, aes(viewed_ninetyfive_percent,title)) + geom_violin()  +
↪  geom_text(aes(label = video_duration, size = NULL, ),check_overlap = TRUE, nudge_x =
↪  0.2)
```



Insights

This chart presents viewer engagement (at least 95% of video) across videos on the MOOC.

- Although *Privacy online and offline* had a high click-rate, it had low engagement.
- Generally speaking, videos with shorter duration had higher engagement.

4. Communicate

Ultimately communication is the most important aspect of exploratory data analysis. Without the ability to present a story to stakeholders, the impact of data analysis will not be felt. This communication imperative applies to technical specialists and non-specialists - either in spoken, written forms and code.

Reporting

This analysis has been documented and reported via R Markdown document, along with a Git log and cache for reproducibility.

Presenting

This analysis has been presented by video over 5 minutes for non-data users.

Future Work

For future analysis of this dataset, more work could be done to explore the correlation between learner variables and their likelihood to leave the course early. This may not infer causality but certain trends may become apparent.

Lessons Learned

The author has learned a lot during this exploratory data analysis of the *FutureLearn Cybersecurity Dataset*, including but not limited to:

- Using R for the first time;
- Using Git for the first time;
- Using plotting tools extensively;
- Manipulating data between dataframes and working in data subsets; and,
- Using a methodology to guide exploratory data analysis.

There are several takeaways, and things to improve for similar kind of work, which are:

- When working to a deadline, try to do little and often rather than a lot in a shorter period - the author spent a lot of time debugging code into the **wee hours**;
- Formatting issues in R Markdown can be very time consuming - indeed there are imperfections in this report;
- A greater exposure to plotting types and statistical methods would have made some of the analysis more straightforward;
- Do not use spaces in directory names; and,
- Data analysis, when approached in a structured manner, can be a lot of fun!

Acknowledgements

The author would like to acknowledge all the support of the academic team whether during a lab session or at 11 o'clock at night. There has been a lot of learning on the job and stumbling through activities but it would have been impenetrable without the careful planning of the assignment and copious ad-hoc support. In particular, I would like to acknowledge:

- Dr Matthew Forshaw
- Dr Joe Matthews