CSC8631 Assignment Report

Marc Birkett

07/11/2021

CSC 8631 - Data Investigation with Student Data

Introduction

Report into investigation of Student Data using the CRISP-DM model. This report covers two iterations of the model and includes the processes of Business Understanding, Data Understanding, Data Preparation. The subprocesses I've chosen are to do the following steps:

- Import
- Tidy
- Visualise
- Understand
- Communicate

The project has been set up using ProjectTemplate to provide structure and repeatability, which will be tested on a regular basis. Version control is provided by Git and this report created with R Markdown. Various libraries have been imported and used from the Tidyverse regarding data import and management (Dplyr, Readr) and visualisation (GGPlot2)

Iteration 1

Iteration 1 was be used to investigate the data and generate a hypothesis for further investigation going through the steps outlined above. Once a hypothesis has been identified this will be further investigated in iteration 2. An initial investigation into each relation in the dataset will be carried out and a graphical summary will be presented with potential hypothesis and further analysis will be carried out on iteration 2. Once we have a working hypothesis summary statistics of the specific data to be used will be presented.

1 - Import and Tidy

The data was supplied in csv files covering 8 different areas of the software over multiple stages, this was imported into R into 8 data frames from the original for easier analysis, and the number of the file was included to give an indication of which stage of the course the data was created which will potentially provide an indication of time for further analysis. After initial investigation is was found that the detected data types were not consistent the following data manipulation was carried out.

- ID - to integer
- Date time - to the local POSIX date time format

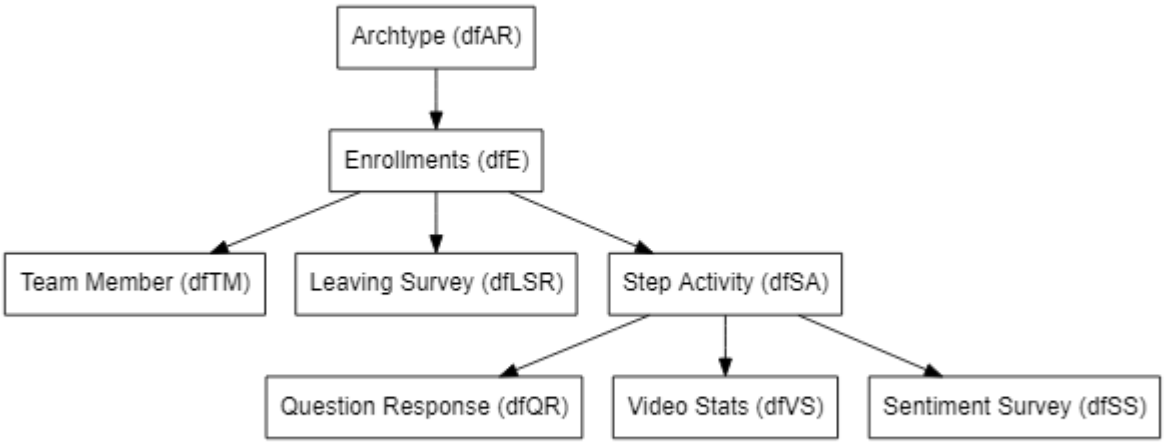


Figure 1: Entity Relationship Diagram showing how to relate each dataframe together

Initial exploration yielded the following entity relationship diagram:

Each relation was also investigated for potential primary keys, this indicated that Learner_Id was a candidate for relating 6 of the 8 data frames as it was a common column, the data regarding video views was video centric and did not have learner id nor did the sentiment survey data. Any data didn't have one had one created using the row_id to uniquely identify the row.

Further analysis during visualisation of the data yielded a relationship between Step Activity, Leaving Survey Responses and Question Response. The "step" data item is a compound of "week_number" and "step_number", which is present in the Step Activity table, this matches "last_completed_step" in Leaving Survey Responses at 1 decimal place. This is also present in Question Response where "quiz_question" is a concatenation of "week_number", "step_number" and "question_number", so we could potentially relate those three relations as an avenue for investigation. The Step Activity relation can then be used to relate to the wider data using Learner_ID

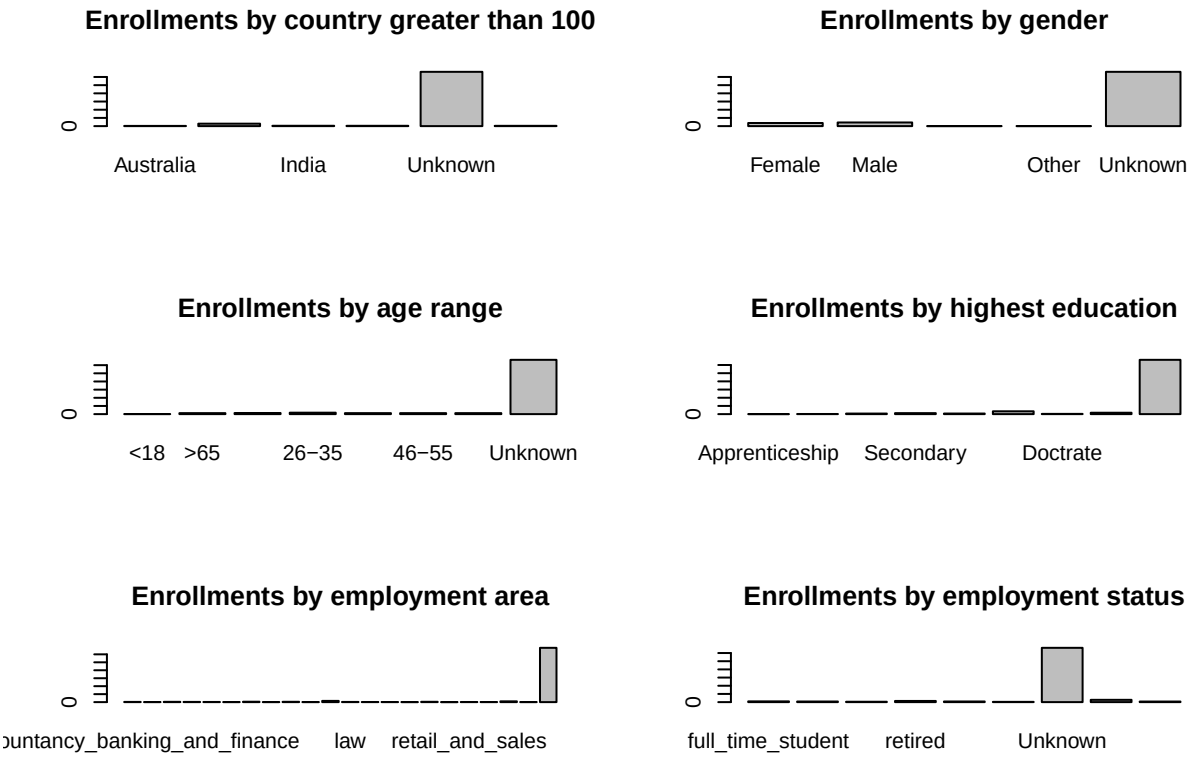
Upon investigation of the data it was decided that further investigation of Enrollments, Step Activity, Leaving Survey Responses, Question Response, Video Stats and Weekly Sentiment Surveys may yield an interesting topic for investigation. The other data was dismissed due to its limited breath.

2 Visualise

Each data frame was investigated with a combination of viewing the data, frequency tables and graphical analysis. Each table identified above will be visualised and some conclusions regarding potential investigation drawn. A comment will be provided on the contents of each table in terms or items, variables and variable types, namely continuous and categorical. Appropriate visualisations for the variable types will be selected. Also an outline of any feature engineering (i.e. the derivation of new data items) carried out will be included.

Enrollments Enrollments is a multivariate dataset with n=37296 items and p=14 variables. This table is principally categorical.

For each variable bar plots were used to visualise the spread of values of the categorical data, these were Country, Gender, Age Range, Highest Education Achieved, Employment Area and Employment Status.

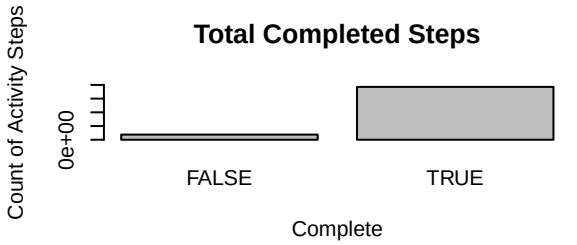


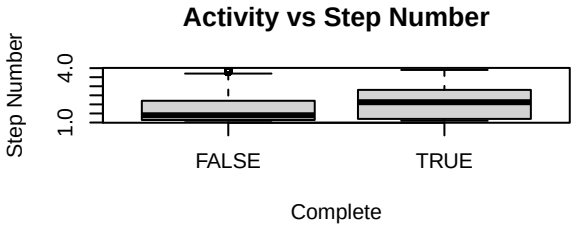
It was identified that the majority of records for each case was "Unknown".

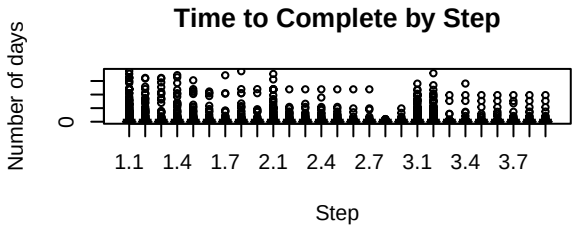
It was concluded the Enrollment data may be used to add richness to an overall data investigation but there was no specific hypothesis to investigate. Potentially this is an avenue to improve data completeness in future enrollments.

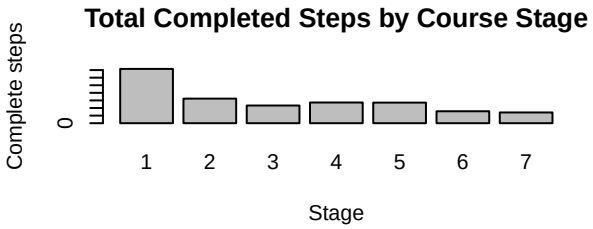
Step Activity The step data is principally categorical data with n=423072 items and p=7 variables, who more were subsequently added after feature engineeering.

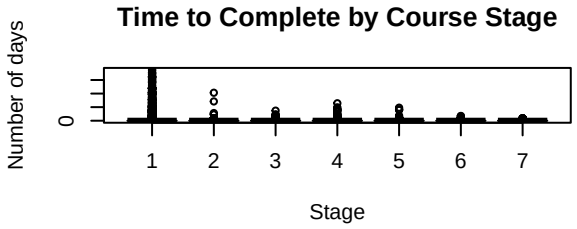
The data includes the step number as discussed at the import and tidy stage and the step start and end date with the learner_id. The end date has been used to indicate whether a step was complete for that student and how long the step took, this has resulted in 2 new columns - "isComplete" flag and "completedTime" in days, and allows us to calculate step completion stats and time scales per step as per below.











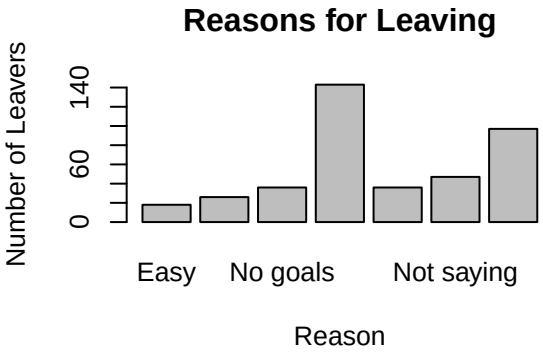
We can see that there are far more completed steps than incomplete, and that incomplete steps are generally earlier in the course. However, there are still students who complete early steps and subsequently fail to complete later ones. There also indicates a wide spread of time to complete each step, with a variance of 103.1396. An outlier of this is *Step 2.8* which takes everyone very few days to complete.

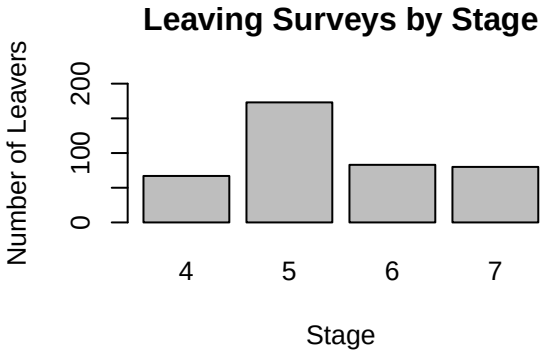
Due to adding the stage of the course we can also see that the time to complete each unit reduces as the course wears on, as does the number of students that complete each step.

The addition of the feature engineered values gives us the potential to use this dataset in further analysis. There is potential to look at incomplete steps in relation to course leavers, and investigate the short completion times of step 2.8.

Leaving Survey Responses The leaving survey responses comprises of categorical, character and date based data with n=403 items and p=10 variables. It is a smaller dataset then previously investigated. As discussed the "last completed step" value of the Leaving Survey data set matches back to the Step Activity data above so this provides the potential to investigate these data set simultaneously if required. The table is predominantly a categorical collection of the reasons for the leaving, and the last completed step and week.

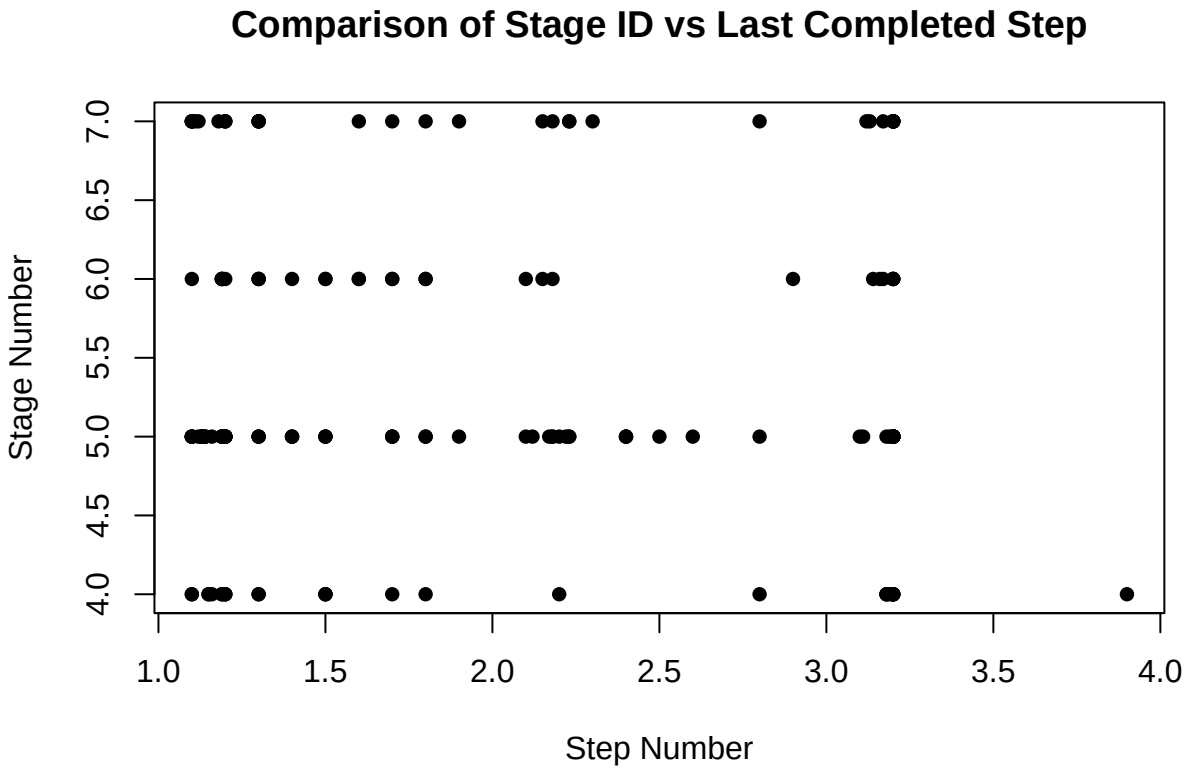
Upon investigation the reason for leaving had multiple categories referring to the lack of time so all reasons that mentioned time were grouped as one to make it comparable with the other reasons.





We can see that the last step completed by the majority of leavers was between 1 and 1.5 with another spike between 3 and 3.5, and also that leavers in the later stages of the course, 4,5,6 and 7 with none prior to that. The majority of reasons for leaving were due to the lack of time. There is the potential to investigate both the time of leaving and the reasons.

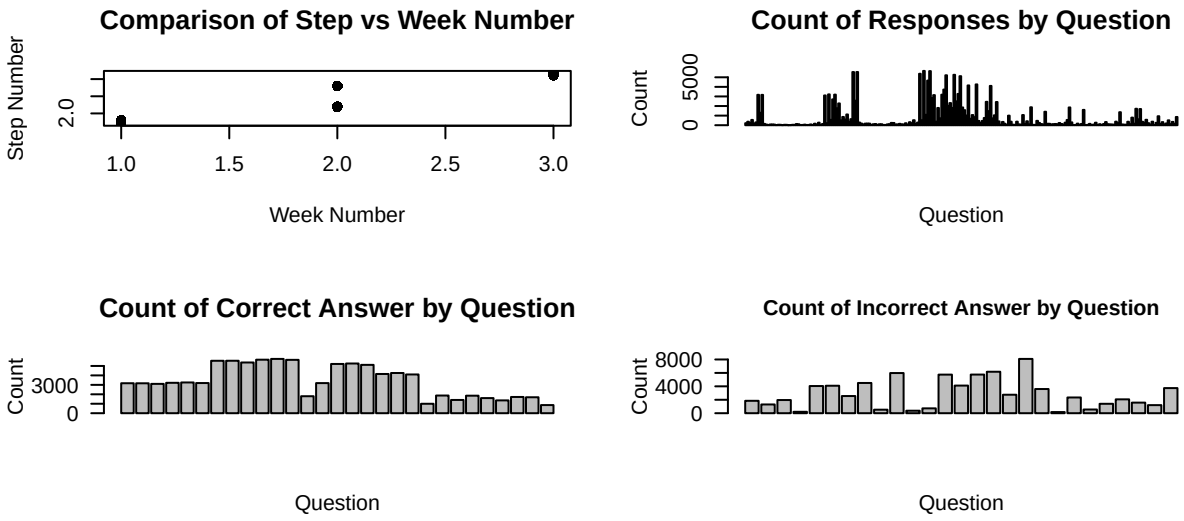
The graphs to compare last completed step and stage seem to be at odd with each other. If students on leaving later in the data set, i.e by stage 5, how come the majority leave at step 1 to 1.5.



As we can see above it seems some leavers leave at the later stages of the course having never completed the earlier steps.

Question Responses The question responses data is categorical data with n=176463 items and p=12 variables after the feature engineeering outlined below. The question responses table has the quiz question which includes the step number. Feature engineering was carried out to extract the correct step number and add it to a column, should there be a need to relate to the the activity steps.

An examination of the data was carried out to look at the volume of responses and outcomes.



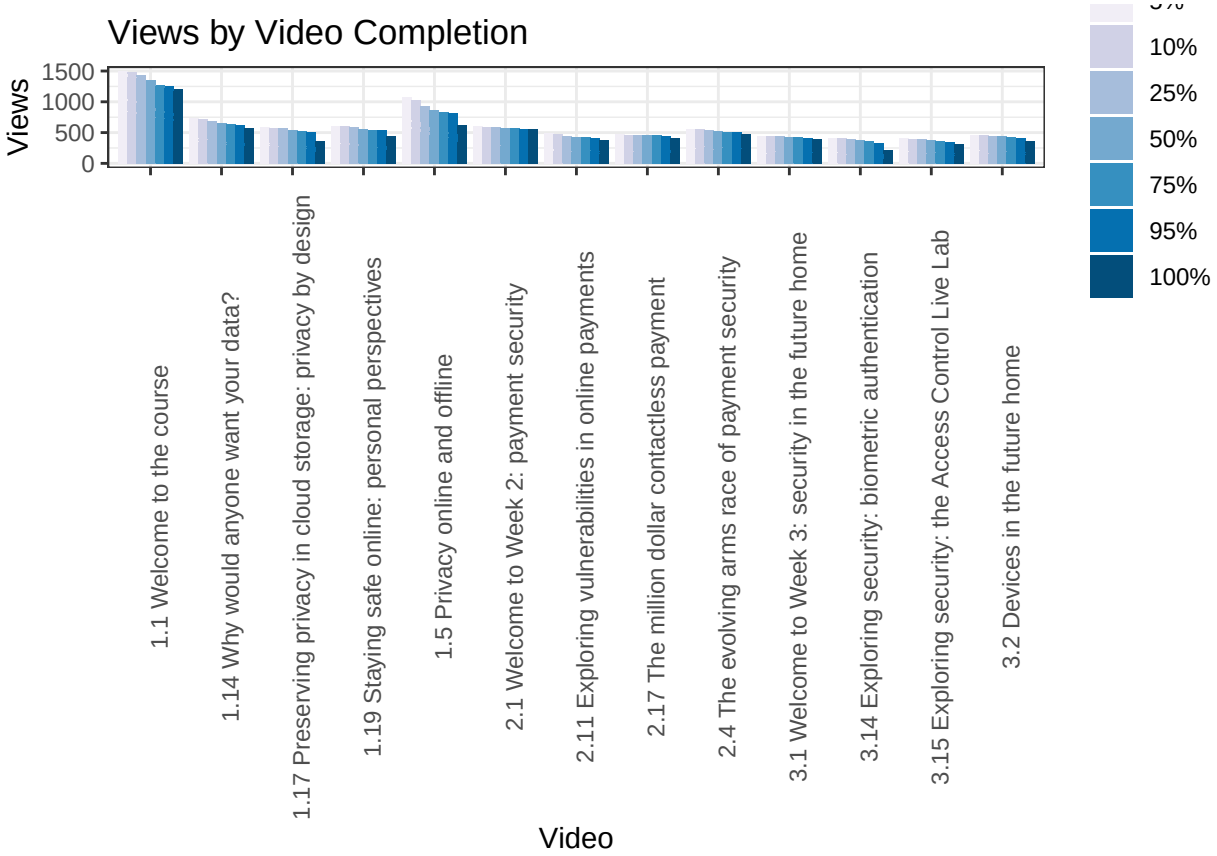
We can see that step and week number generally match which indicates that step is related to week throughout the course. This also indicates that what I've called "stage" also indicates step and week number.

As we can see there are significant differences in number of times each question was answered over the length of the course and the as expected mix of correct and incorrect answers. There is the potential here to look into the mix of correct and incorrect answers, potentially investigating any relationship between low attainment on the quizzes throughout the course to the student subsequently leaving.

Video Stats The video stats data is multivariate data with n=65 items and p=29 variables. There are significantly more data variables in this relation then in those examined previously, and significantly less data items.

The data in this relation can be subsetted into data pertaining to percentage complete, type of device and location of the video view, so three data sets were created with just this data present. Some work was then done to convert the data from percentages to raw numbers, and then to pivot the result to allow some insights into the number of views.

Views Percentage Complete It can be seen below that the number of views which watched the whole video dropped off as they viewer got further through. For example we can see for the first video "1.1. Welcome to the course" that 1500 views completed 5% of the video, which drops to 1200 who completed 100%. This is then presented for each video throughout the course.



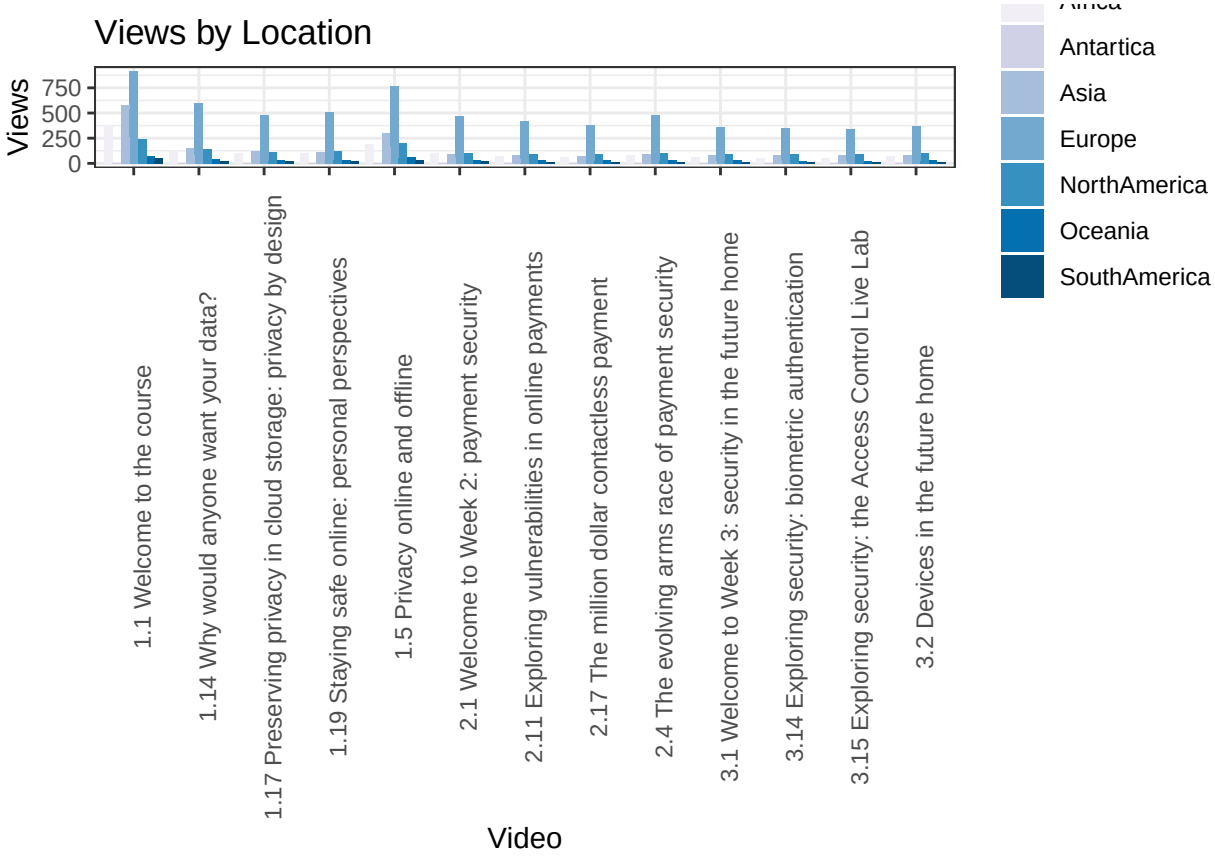
It can be seen that as the course progresses the volume of students engaging with each video reduces reasonably significantly other than "1.5 Privacy online and offline" It can also be seen that some videos see a significant drop between 95% and 100% completion.

Views by Device When viewing views by device it can be seen that no video was viewed on a tablet or TV, all views were by Console, Desktop or Mobile. It can be seen that by far the majority of views were on a Console throughout the length of the course, so it may be concluded the videos should be designed to play best on a console.

```
{ r plotVideoDevice, echo=FALSE}   #video by device   ggplot(data = dfVSDevicePivot,
aes(fill=percentviewed, y = count, x = as.character(title))) +     geom_bar(stat="identity",
position="dodge") +     labs(title= "Views by Device", y="Views", x = "Video") +      theme_bw()
+      scale_fill_brewer(palette="PuBu", name="Device")+     theme(axis.text.x = element_text(angle
= 90))
```

The data shown reflects the same changes in the volume of use as shown above, i.e. that "1.5 Privacy online and offline" shows an increase in views in an otherwise downward trend.

#####Views by Location By viewing views by location it can be seen that the majority of views are from Europe with non at all from Antarctica and few from North America, Oceania and South America.

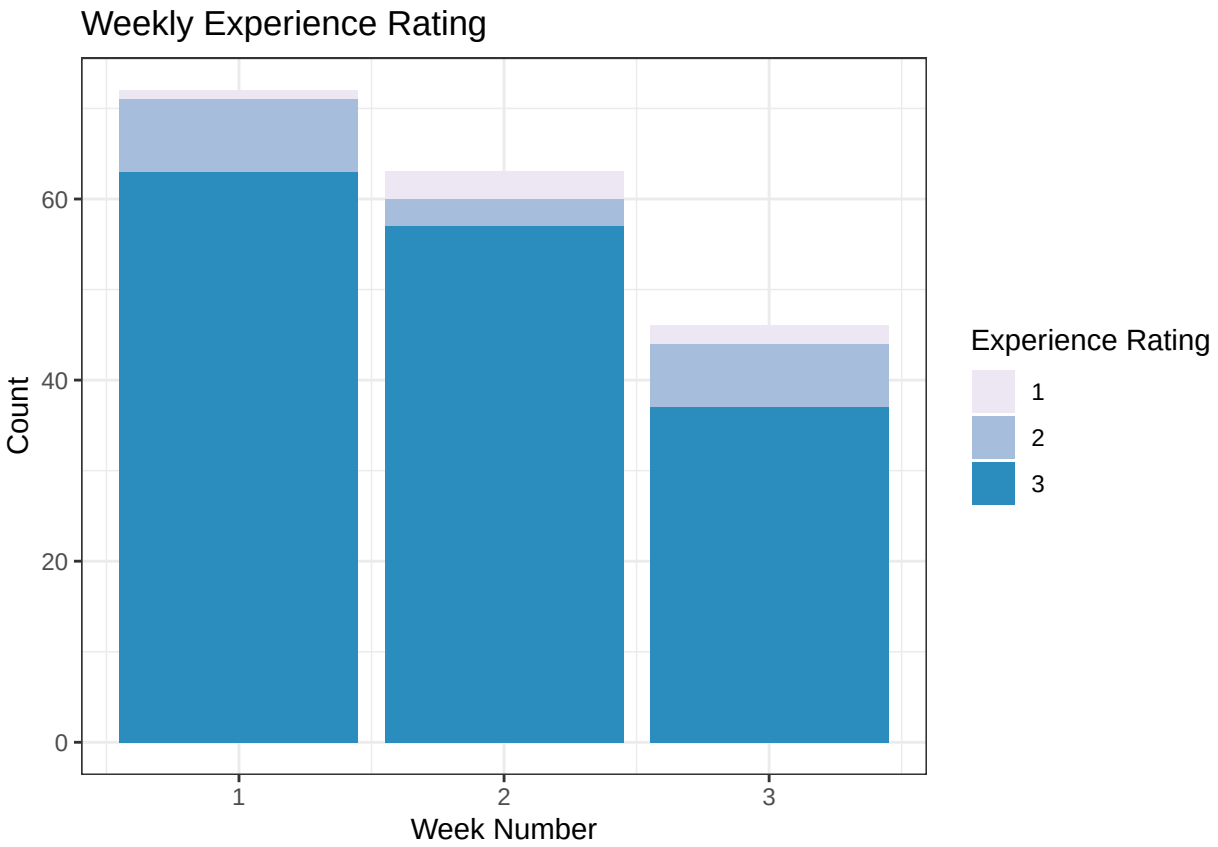


The data shown reflects the same changes in the volume of use as shown above, i.e. that "1.5 Privacy online and offline" shows an increase in views in an otherwise downward trend.

There is potential to link views with stages and we could potentially look at leavers by location and device by using this view data along with the leaver survey responses. We could also look at test performance related to video completion, location or devices.

Weekly Sentiment Analysis The video stats data has n=181 items and p=6 variables. This table has a categorical properly rating and character data to cover the sentiment. I will not be looking to cover sentiment analysis in this work so only the categorical value will be looked at.

As we can see from the weekly experience data below, across the three weeks of the course the total sentiment data provided by the student cohort went down from 73 to 45, although the experience ratings being provided remained consistent at approximately 90% providing a rating of 3.



It would be of value to understand this data as a percentage of the total patient cohort. Given the lack of a "learner_id" in this data it isn't immediately possible to relate the sentiments back to a specific learner, so this data could only be looked into by step. There may be value in understanding after which step students decided to stop providing feedback, this could be related on the "stage_id".

3 Understand - Conclusion and Hypothesis

Step, "Stage" which is the number in the file name potentially the same, needs looking at - do they match?

Potential Hypotheses Test performance vs video views, devices or location Leavers vs test performance Leavers vs location Sentiment vs Step.

Iteration 2 Hyptohesis Test performance vs video views vs leavers

Iteration 2

Iteration 2 will further investigate the hypothesis identified in iteration 1 and will present the findings.

Location Analysis Multivariate Analysis GGPlot2

Findings

To answer the hypothesis XYZ the findings are that ABC

Conclusion

References

ReadR - https://readr.tidyverse.org/index.html DiagrammeR - https://rich-iannone.github.io/DiagrammeR/ Working with categorical data - https://cran.r-project.org/web/packages/vcdExtra/vignettes/vcd-tutorial.pdf

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

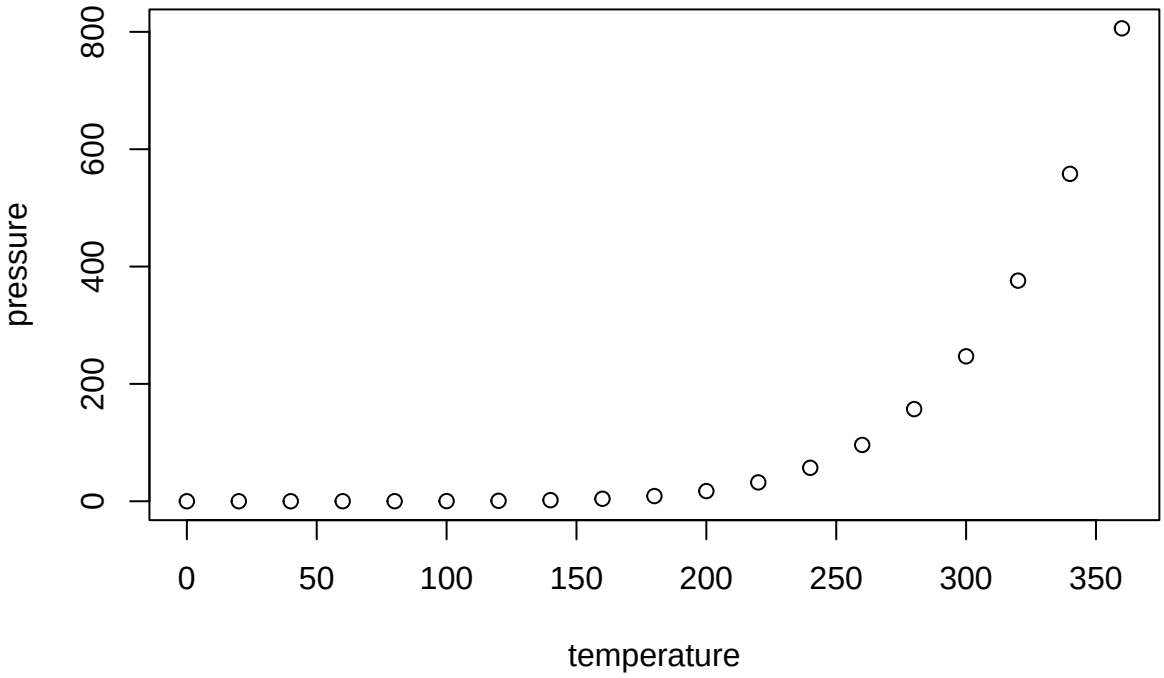
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.