# 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 some structure and repeatability, which will be tested on a regular basis. Version control is provided by Git and this report created with R Markdown.

## 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. A graphical summary of each data set will be presented with potential hypothesis and further analysis will be carried out on iteration 2.

### 1 - Import and Tidy

The data was supplied in csv files covering 8 different areas of the software over multiple stages, this 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 and to potentially aid 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

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

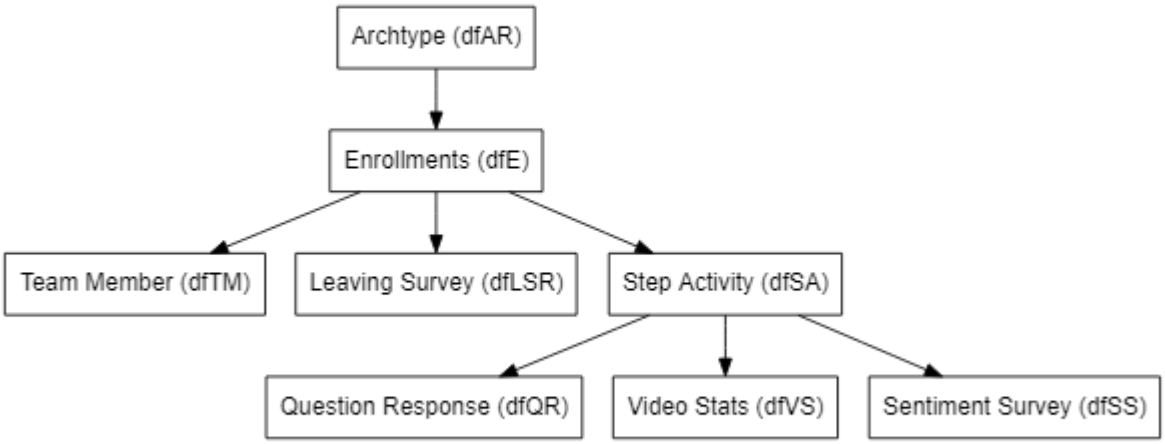



Figure 1: Entity Relationship Diagram showing how to relate each dataframe together

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.

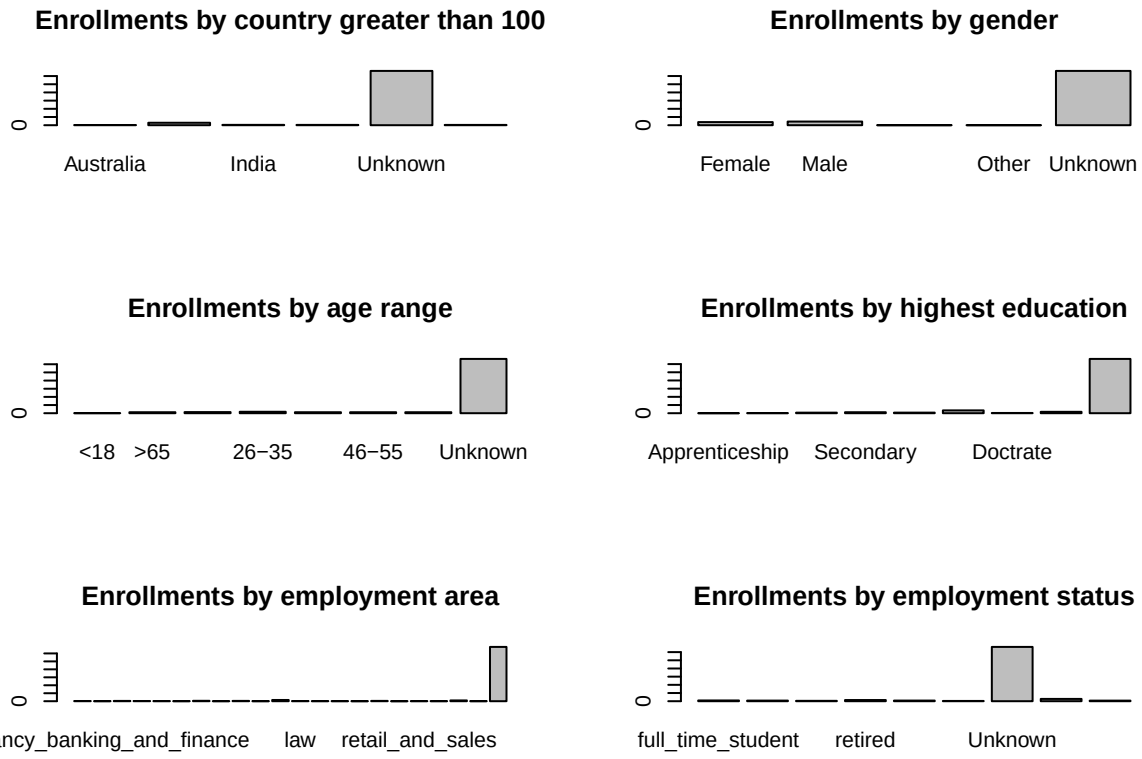
Further analysis during visualisation of the data yielded a relationship between Step Activity, Leaving Survey Responses and Question Response. The "step" data item is a compound of "week_number" and "step_number", which is present in the Step Activity table, this matches "last_completed_step" in Leaving 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.

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

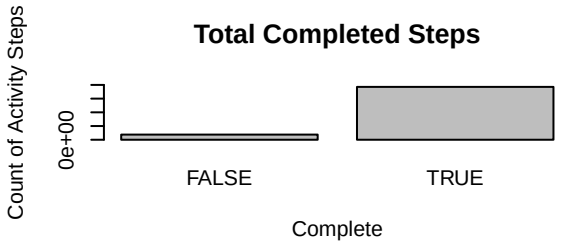
Each data frame was investigated with a combination of viewing the data, creation of barplots, histograms, pie charts and frequency tables. Each table identified above will be visualised and some conclusions regarding potential investigation drawn. Data was related and cast as required.

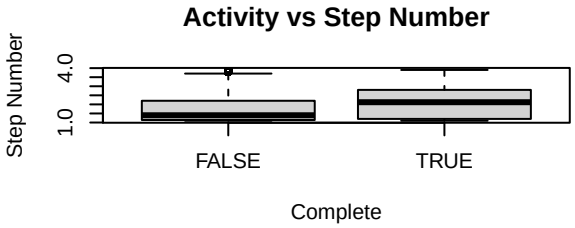
**Enrollments** Enrollments is a multivariate dataset with n=37296 items and p=14 variables. For each variable barplots were used to visualise the spread of values in mainly categorical data. Data items invesigated were Country, Gender, Age Range, Highest Education Achieved, Employment Area and Employment Status which identified that the majority of the data was unknown in each case.

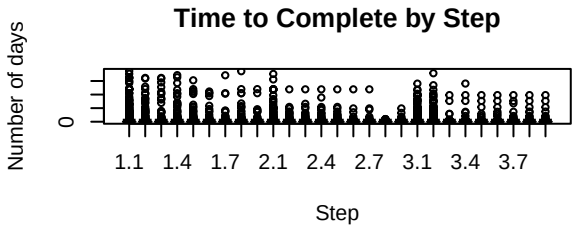


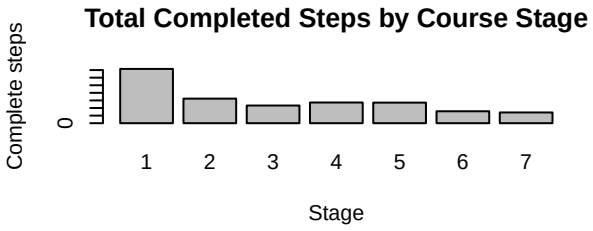

It was concluded the Enrollment data may be used to add richness to an overall data investigation but there was no specific hypothesis to investigate.

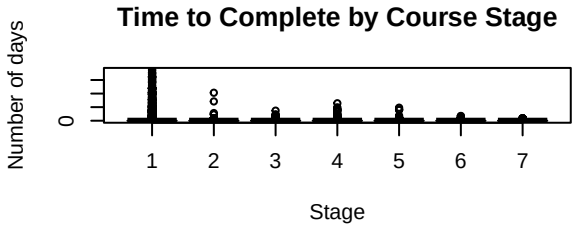
**Step Activity** The step data is multivariate data with n=423072 items and p=9 variables after the feature engineeering outlined below. The step data includes the step number as discussed at the import and tidy stage and the step start and end date with the learner_id. I have used the end date to indicate 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.



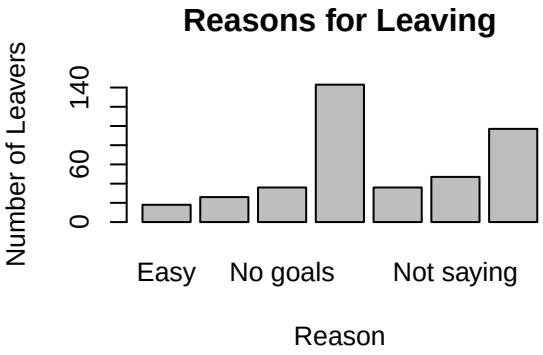

This shows us that there are far more completed steps than incomplete, and that incomplete steps are generally earlier in the course. However, there are still students who complete early steps and subsequently fail to complete later ones. There also indicates a wide spread of time to complete each step,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

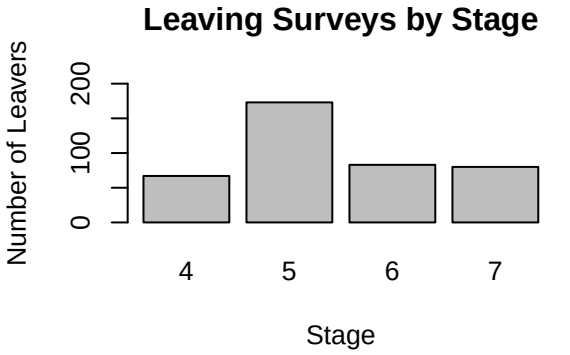
Due to adding the stage of the course we can also see that the time to complete each unit reduces as the course wears on, as does the number of students that complete each step.

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 respones is multivariate data with n=403 items and p=10 variables. It is a smaller dataset then prevously 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 predominantely a categorical collection of the reasons for the leaving, and the last completed step and week. I will use step.

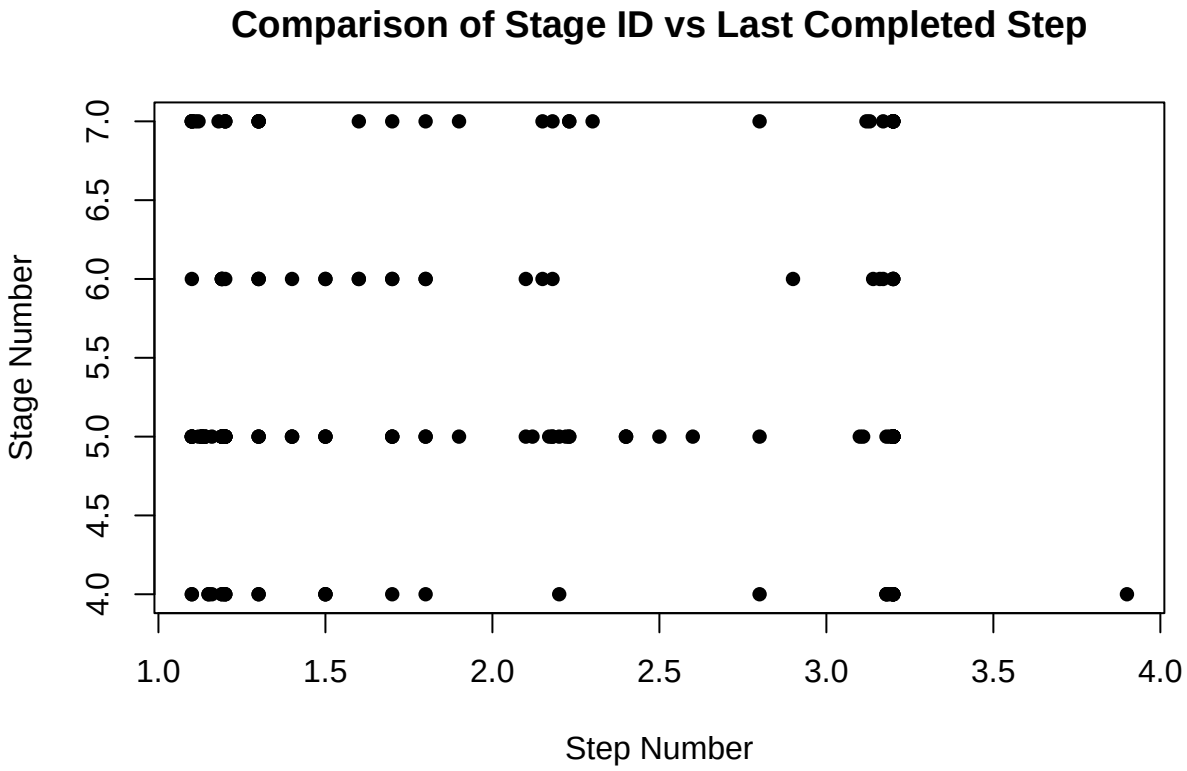
Upon investigation the reason for leaving had multiple categories referring to the lack of time so all reasons that mentioned time were grouped as one to make it comparable with the other reasons


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.

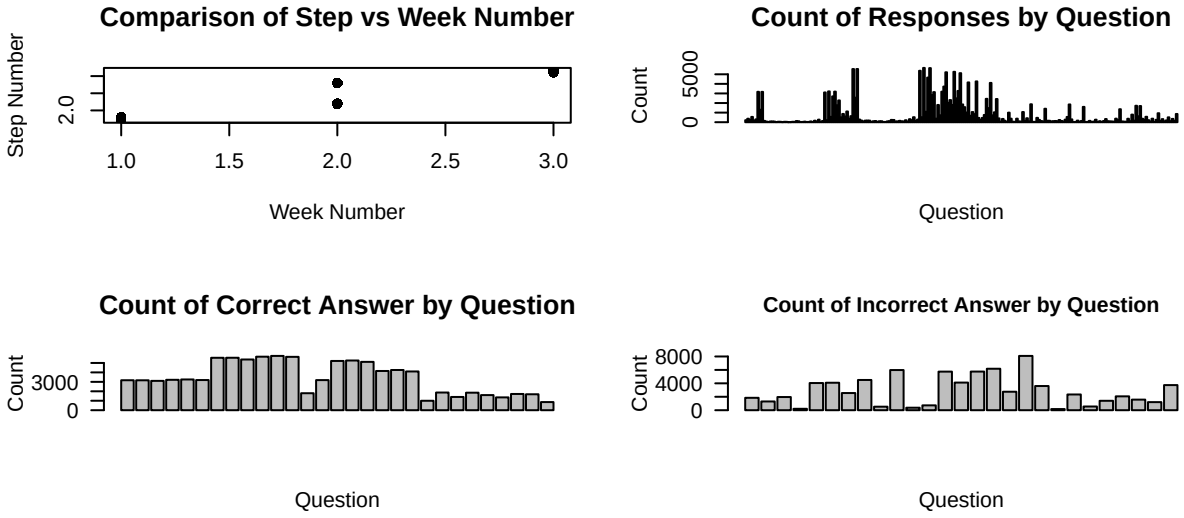
The graphs to compare last completed step and stage seem to be at odd with each other. If students on leaving later in the data set, i.e by stage 5, how come the majority leave at step 1 to 1.5?




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

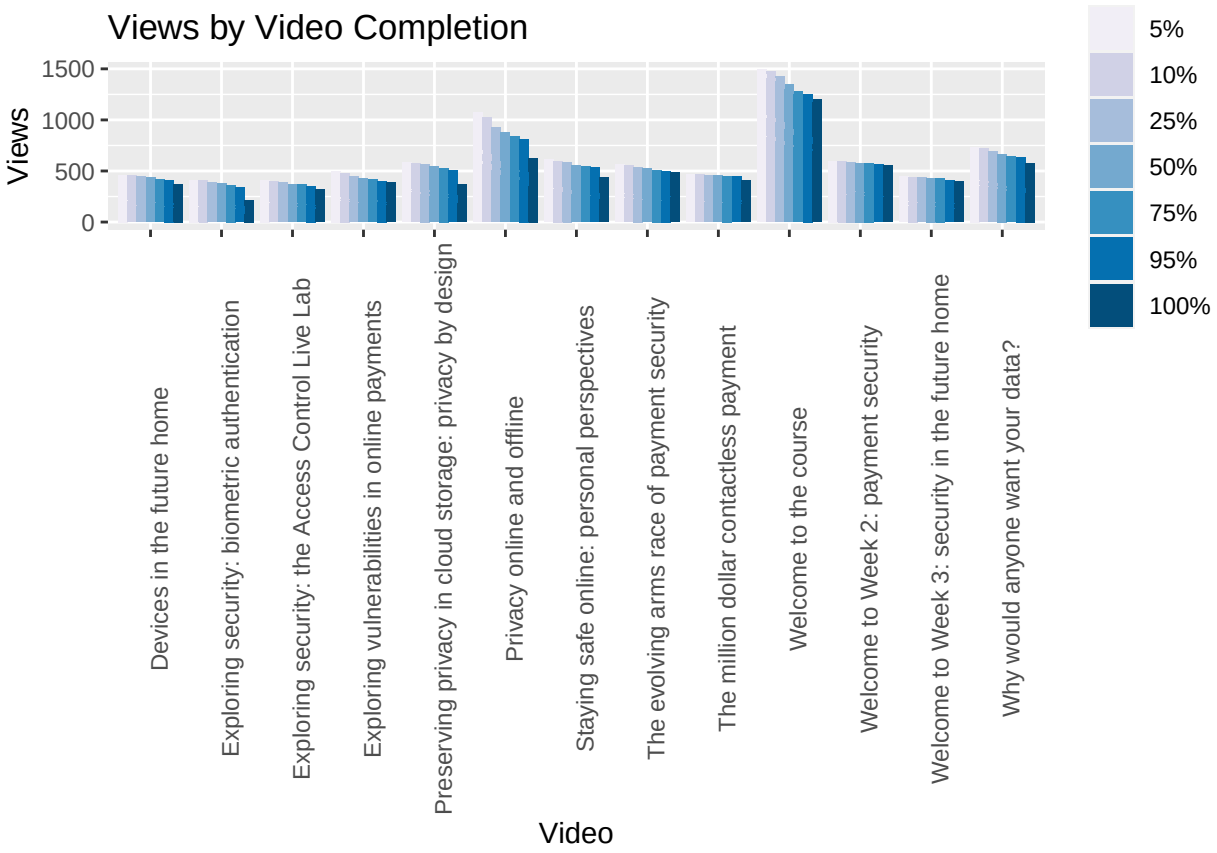
**Question Responses** The question responses data is multivariate data with n=176463 items and p=12 variables after the feature engineeering outlined below. The question responses table has the quiz question which includes the step number. Some feature engineering was carried out to extract the correct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

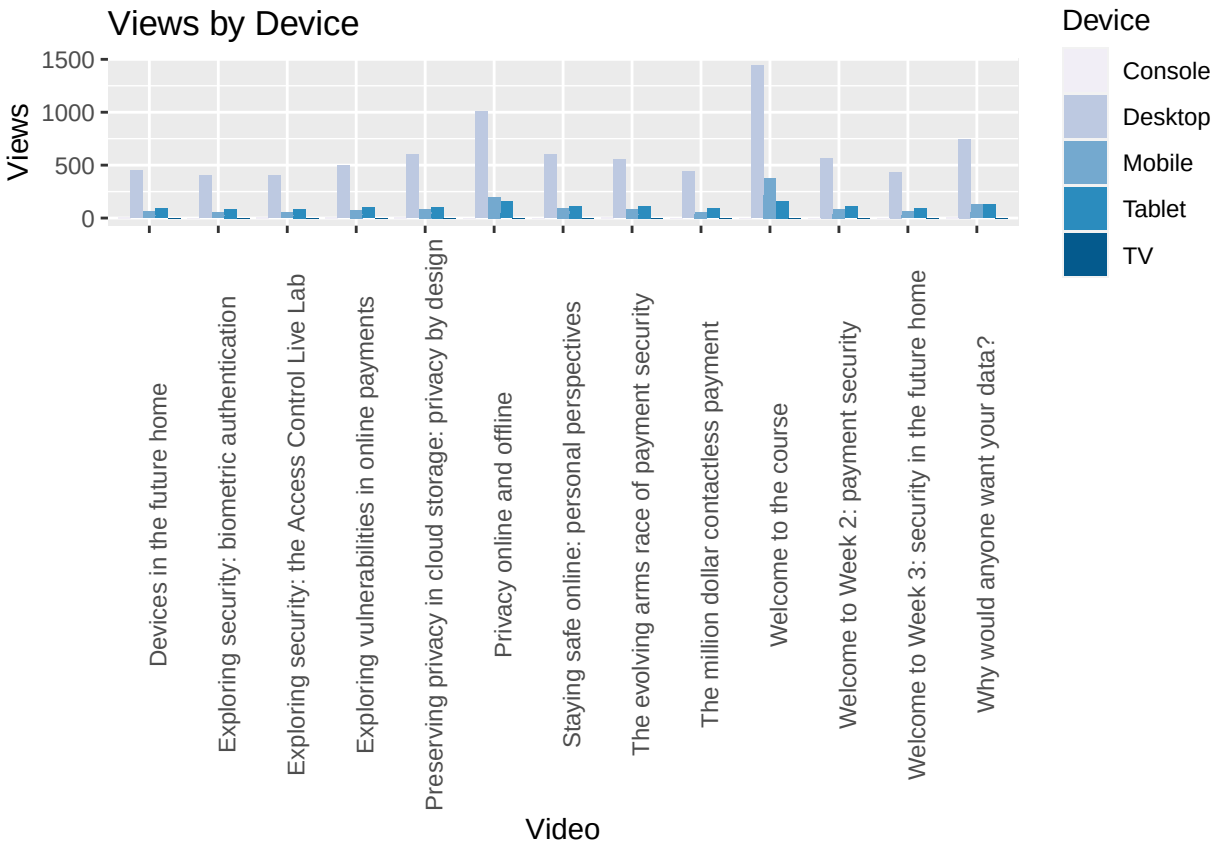
NARRATIVE!

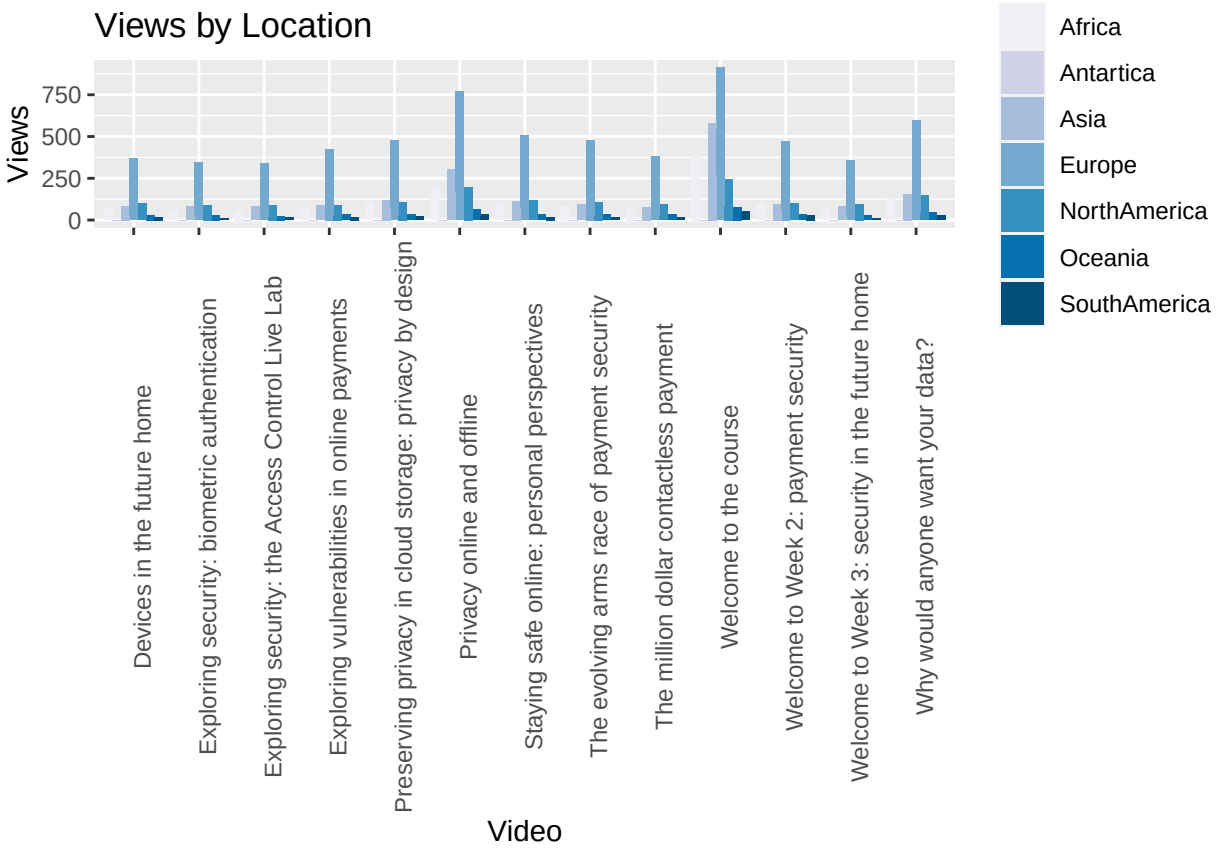



**Video Stats** The video stats data is multivariate data with n=65 items and p=29 variables. There are significantly more data variables in this relation then in those examined previously, and significantly less data items. To handle the increased number of data items, the data was split into views by percentage complete, type of device and location of the view.

NARRATIVE!


Views by Video Completion
Views
1500
1000
500
0
Devices in the future home
Exploring security: biometric authentication
Exploring security: the Access Control Live Lab
Exploring vulnerabilities in online payments
Preserving privacy in cloud storage: privacy by design
Privacy online and offline
Staying safe online: personal perspectives
The evolving arms race of payment security
The million dollar contactless payment
Welcome to the course
Welcome to Week 2: payment security
Welcome to Week 3: security in the future home
Why would anyone want your data?
Video
5%
10%
25%
50%
75%
95%
100%


Views by Device
Views
1500
1000
500
0
Video
Devices in the future home
Exploring security: biometric authentication
Exploring security: the Access Control Live Lab
Exploring vulnerabilities in online payments
Preserving privacy in cloud storage: privacy by design
Privacy online and offline
Staying safe online: personal perspectives
The evolving arms race of payment security
The million dollar contactless payment
Welcome to the course
Welcome to Week 2: payment security
Welcome to Week 3: security in the future home
Why would anyone want your data?
Device
Console
Desktop
Mobile
Tablet
TV




**Weekly Sentiment Analysis** The video stats data is multivariate data with n=181 items and p=6 variables. This table has extensive character based data.

NARRATIVE!

```
## NULL
```

#### 3 Understand - Conclusion and Hypothesis

## 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.

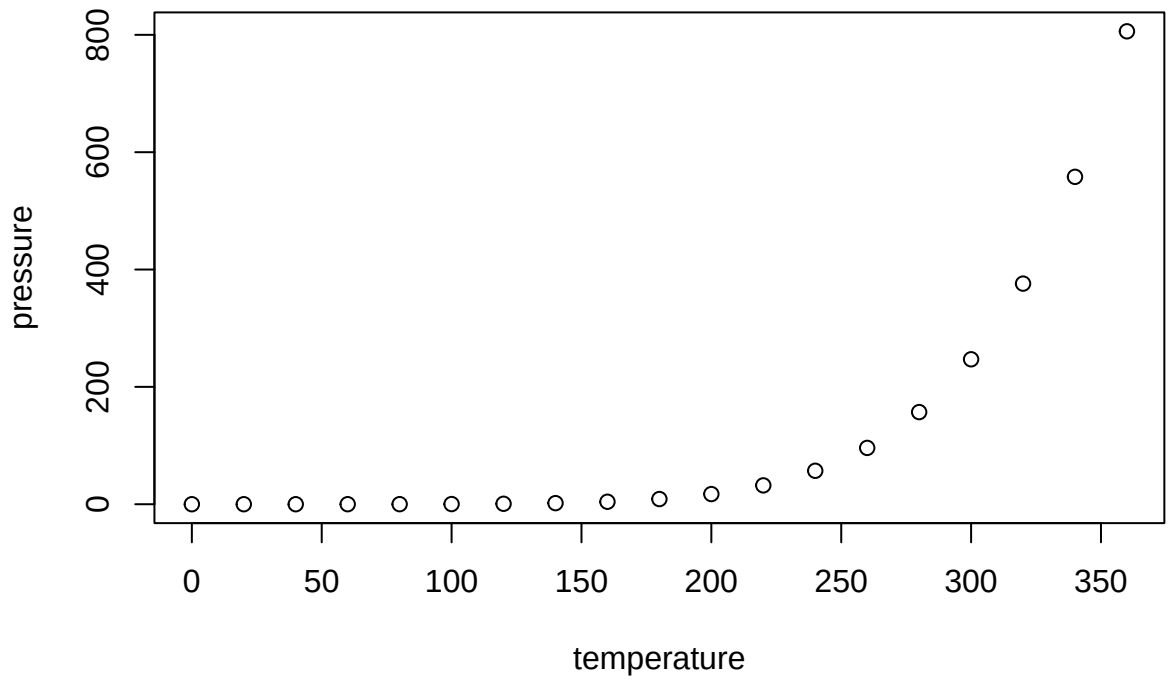
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
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.