Performance-learning analytics

Callum Matthew Simpson

The following is my report of my Data Management and Exploratory Data Analysis course work. It will discuss my analysis and how I did it using the Crisp-DM model.

Business understanding

As the first step in the Crisp-DM model is business understanding I first looked into the task that I had been given and who gave it was for.

The Client background

The client for this report will be Newcastle University who is using the site Future Learn to host a course on Cyber Security: Safety at Home, Online, in Life. Future Learn is an online training provider offering a wide selection of courses, often delivered in association with big-name colleges. Newcastle University is a UK public research university based in Newcastle upon Tyne.

The objective

Newcastle university has gathered information about the users that have attended this course and are wanting to perform learning Analytics on this data to get insight into student engagement.

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

As there is no formal requirement on the question that the university wants to be investigated I decided that I wanted to get a better understanding of why a user leaves the course. I feel like this would be a good measure of engagement as If could the common reasons and trends that are causing people to want to leave the course, then with these findings future courses improve in ways that would stop people leaving the, keeping users engaged and lower the amount of people that leave the course.

Because of this my business object will be "give useful insights into why people decide to leave the course"

Assess situation

I have been given data about the users who have attended the course, like enrollment data and answers to activity question. The data is split across seven run of the course (so each run has it own set of user data).

Determine data mining goals

As I'm wanting to look into useful insights into why people decide to leave the course a clear data mining goal will be finding out the common reason a person leaves so that the course could be updated to help addresses these issues to reduce people leaving in the next cycle. Similarly another goal of mine should be

be finding out who is leaving as again it means that measure could be put in check to increase engagement in that group. My success criteria will be data that clearly allows us to see why people are leaving.

I will deem my analysis successful if the data I use to come to the conclusion is clean (or issues can be explained)

Project plan

My aim is to

- understand data by the 17th of November
- have the first clean bit of data by the 22nd of November
- Have roughly two weeks to do my analysis
- get my analysis done by the 1st of December.

I am planning to use and R project with a project template to ensure that my code can be reproducible.

Data understanding

The Collect initial data

The data I have been given is a set of different datasets relating to a different factor of the course. We have 7 of these "set" as we information about 7 runs of this course. New modifed data based off / using this data will be created automatically by my project template. I will say when its relevant when this has been done.

Description of all data

The following will go over all the data collected and briefly outlines what is stored in each dataset. As all the data has been made my the one source I will assume that all the data can be merged together using common meta data without any issue.

Archetype survey responses Only data for cycle 3 to 7 seems to be recorded. It a collection of the archetypes of the users with 4 columns.

- id - user id for cycle
- learner_id - user id
- responded_at - the time a user responded
- archetype - what Archetype they define themselves as

Cyber security enrolments For all cycles,The enrollment data for each user. Contains 13 rows. Note worth rows are

- id - user id for cycle
- learner_id - user id
- role - role of the person one the course
- gender - User identifed gender
- country - give users country
- age - ages done in age range brackets
- highest education - Highest education.
- It also contianes if they are employed and where

cyber leaving results Only for cycles 4 to 7. The reason people left and when about in the course they left.

- id - user id for cycle
- learner_id - user id
- left_at - time and date when a user left the course
- leaving_reason - the reason why the user left given from a set selection of given options
- last_completed_step_at - time and date of the users last step
- last_completed_step - what was the users last step
- last_completed_week_number - what was the last week a user got to
- last_completed_step_number - what was the users last step the completed

question response For all cycles. The responses for the questions given throughout the course.

- learner_id - user id
- quiz_question - the full question numner
- question_type - the type of question
- week_number - the week number the question appread
- step_number - the step the question apperead
- question_number - the question sub number
- response - users answers
- cloze_response - answer approximated
- submitted_at - The time a user submitted the question
- correct - Was the user correct.

Step activity For all cycles. When a user started and finished that step

- learner_id - user id
- step - week + step number
- week_number - the week number
- step_number - the step number
- first_visited_at - When the user first visited the step
- last_completed_at - when the user finished the step

Team members only in cycles 2 to 7. Contains information about peoples rows.

- id - users id
- first_name - users first name
- last_name - users last name
- team_role - team role is the role whole of the person in the team
- user_role - user role seems to be role of what the user

weekly_sentiment_survey_responses Seems to be only cycles 6 and 7. A brief survey about how people are finding the course

- id - user ID
- responded_at - the time the user responded.
- week_number - the week
- experience_rating - out of 5 how did the user rate the week
- reason - a comment on their opinion

video_stats For all cycles. Information about the videos. There are 23 columns so instead of listing everything out this data set contianes

- how long people are watching a video
- where they are watching it
- what people are watching it on

Overview

As you can see we have allot of data at hand to help us answer the question. Each dataset has an ID function so that can be our meta data that can be used to link all the dataset together. Exploring the dataset its clear that not all cycles contain the same recorded data as others. This has lead me to make my first assumption.

Assumption One

Not all data was recorded from the start. I am assuming that this is because as the cycles went on new aspects user interaction where decided to be recorded.

Choosing the right data

After doing a scan over all the data I decided on what I was going to use. As I decided to look into why user left the course and obvious start would be by looking at the the courses leaving results to see what users where saying.

As I had 4 datasets to work with (cycles 4 to 7) I merged them together but made sure that I included an new column called Cycle which contains the cycle number that that row of data came from. This made a dataframe of 403 rows.

The following section will be using this dataset.

Explore Data and explaining

ID variables

As I merged 4 datasets together I wanted to make sure that there was not any weird duplicates or any in cycle duplication for our 2 variables are relating to an id.

The generic ID

```
##
## FALSE
##   403
```

The user ID

```
##
## FALSE  TRUE
##   370    33
```

We see that id has no duplicates but learner_id did contain some duplicate ids.

From some extra checks we see that id is a unique ID for the user on each cycle and user ID is the ID of the user.

From a check for duplicates leaner Ids we find that for the most part are the same learner id but in a different cycle suggesting that we have a few people who are leaving in one cycle but rejoining another cycle only to leave again later. This is fine as its same user in different cycles and not the same user in same cycle.

If we look at duplicates of learner_id and cycle we find 15 occurrences. For the most part we find alot of these duplicates are within a few minuets of each other, However some of these dates are days out.

To be safe I decided to remove all these duplicate values leaving us with 388 rows to work with.

How many people left each cycle

Next I wanted to see how many people left in each cycle so I create a table of counts per cycle.

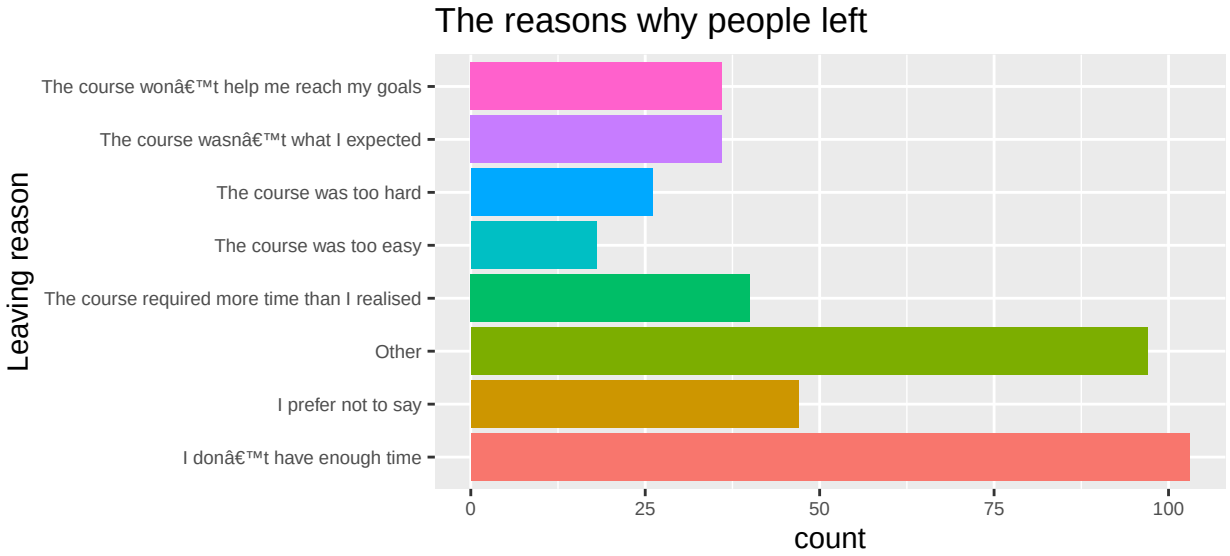
```
##
##   4   5   6   7
##  67 173  83  80
```

As there is no NA colunm we can safely ensure that every row has be assigned to a cycle.

As we can see there was almost double the amount of people leaving in cycle 5 as compared to other cycles suggesting that either + a) Alot more of people joined in cycle 5 than compared to the other years meaning the same percentage of people left. + b) Something caused alot of people to leave this year than compared to the other years

Reason for leaving

As we can see the reason why people leave the course we can try to find the which was the main reason for people leaving.



Looking at all the leaver results put together we see that the majority of the user left as they don't have enough time or for another unknown reason (other). We see that there is a few more people saying that they

left because the course was too difficult than there is saying the course is too easy, this suggest that one the whole people seem to find the course hard than easier.

We also see that there inst any NA values suggesting that the data is complete and every row contains a reason for leaving.

last_completed_step The last completed step tells us how far a user made it through the course. I expect that the majority of the people leave at the start. last_completed_step has a lot of different factors so visualizing the results may be difficult.

```
##
##  1.1 1.11 1.12 1.13 1.14 1.15 1.16 1.18 1.19  1.2  1.3  1.4  1.5  1.6  1.7  1.8
##   12    1    2    6    3    2    2    1   14   15   12    4    8    3    7    6
##  1.9  2.1 2.12 2.15 2.17 2.18  2.2 2.22 2.23  2.3  2.4  2.5  2.6  2.8  2.9  3.1
##    2    2    1    2    1    4    2    1    4    1    3    1    1    3    1    1
## 3.11 3.12 3.13 3.14 3.16 3.17 3.18 3.19  3.2  3.9 <NA>
##    1    1    1    1    1    2    4    2   44    1  217
```

There are 32 individual steps on the course and though out the database there is some steps that people don't leave at. Looking at the table of results we see that there the is alot of (over 50%) NA values in this data set suggesting that the quality may not be that great. Whilst we see a few people leaving at the start that slowly decrease over time (though it picks up near the end of week 1) we see an stark increase at step 3.2 which suggest that there is something at that step that is causing people to leave.

last_completed_week_number I decide to keep the last week number in as it would be useful to group all of the week steps together so we can record how far a user got through the course by large milestones and not small steps.

```
##
##    1    2    3 <NA>
##  100   27   59  217
```

Looking at the table of results we see that there the is alot of NA values(over 50%) in this data set. What is interesting is that there are the same amount of NA values that where found in the last_completed_step suggesting that there might be some similarities in these NA values which might be worth looking into to.

Not looking at the NA we see that the first week has alot of people leaving and not that many leaving in the second week.

last_completed_step_number For the the rest of my analysis I will likely ignore this column as it is the step number of each week. As we have 3 weeks of content then we will have duplicate step number values for each step so its will be likely that is will cause confusion down the line. The last_completed_step contains the week and step number so it wouldn't have this issue. Because of that I will not need this column.

Left at The last column to check was the leave date. A quick check of this column showed that there was no error and each user had a leaving time.

Data preparation

Dataset

The data set that I will use is the dataset of the combined leavers responses. As we saw in the data exploration we have some issues with the data that I will need to explore and try to fix. The following section the will be exploring this dataset and cleaning it.

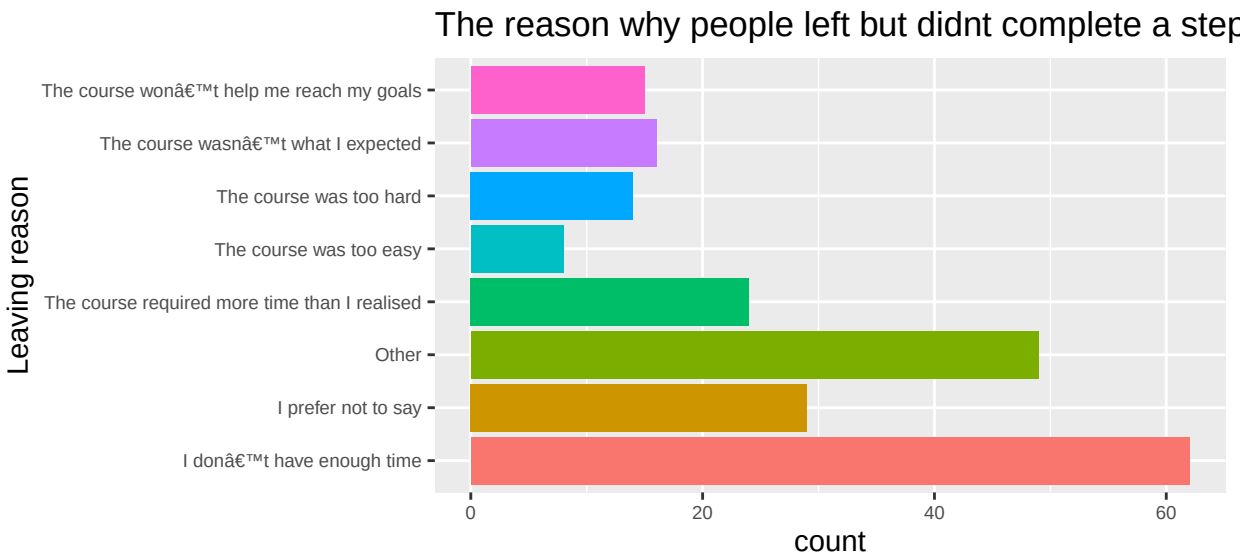
The NA rows in the last_completed_step and last_completed_week_number As the last_completed_step and last_completed_week_number are more meant to represent week titles I decided to convert them from doubles to factors so that I could better utilize the data. I also did the same to the column cycle.

When looking the last_completed_step and the last_completed_week_number columns they both contained 204 Na rows. As they both are the same value this suggest that if one column is NA then the other will be NA. Doing a check on this got me.

```
## [1] 217   9
```

When we get all the rows that contained NA we have 217 Which tells us if a record has NA for last_completed_step there is also gonna be NA for last_completed_week_number and last_completed_step_number. As 217 rows is over 50% of the data I wanted to think of a reason why this may be the case.

My assumption is that either - something has wrong with data collection - These people have left the course before they had done anything on the course so a week value wasn't recorded



If we plot out the reasons why these NA people left we find out that the majority left due to them not having enough time or for other reasons. If these people NA people are infact leaving at the start of the course this results sort of makes sense (they joined the course but discovered they didn't have much time to work on it so they to leave at the start). We see that alot more people say the course was more difficult than it was easy suggesting that they may be a difficulty curve at the start of the course may be too high for people to eneter.

Assumption two

All NA values in the last_completed_step and last_completed_week_number column are actually week "zero" (didn't do anything on the course). From now on all of these NA values will be converted to Zero.

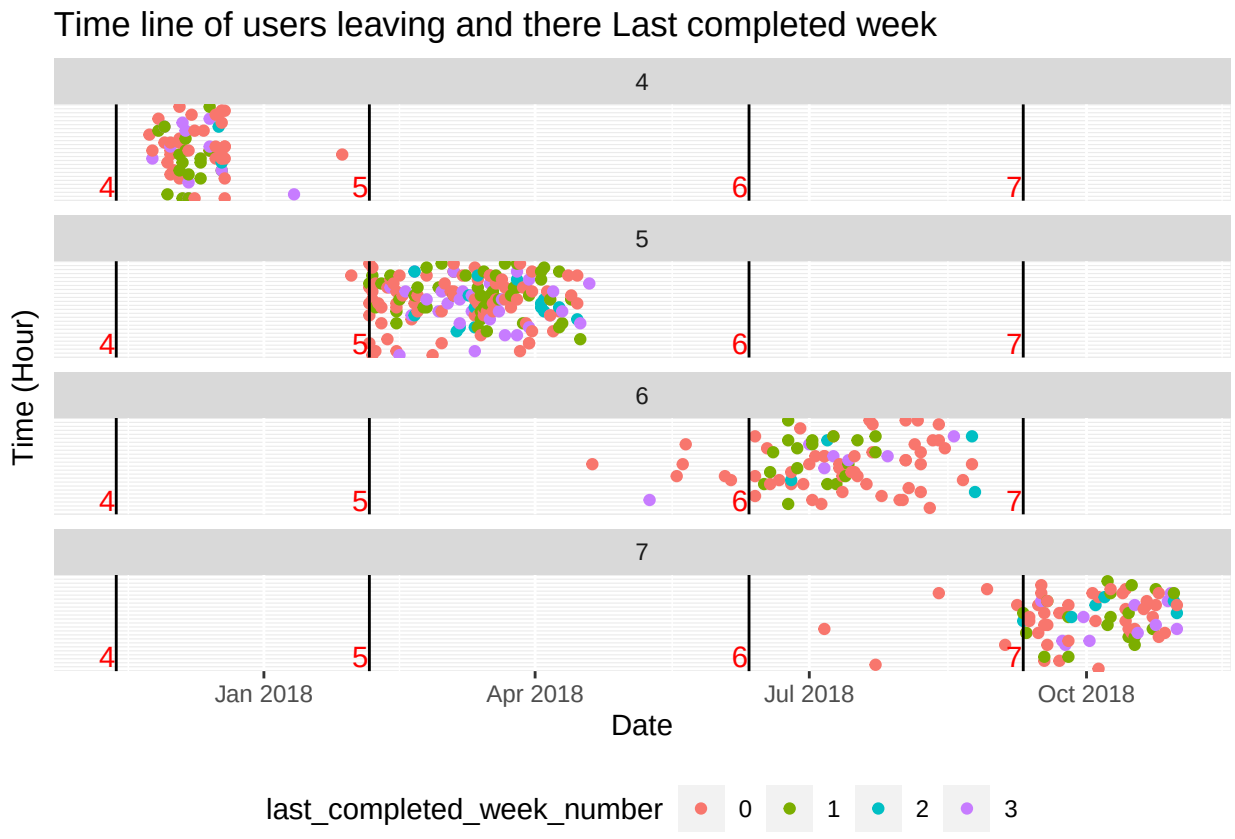
The day people left Another thing i wanted to look was left_at, This is moment in time where a person left the course.

For this I had to do some work on the database. I have changed left_at to a actual time step and then split it onto two new columns Date and time. This was so I could create an upcoming plot that I believe would be useful. The left_at value is very spefic moment in time when people left so to really understand the moments people left we are going to need to split the result into there original cycles (as they leave one after the another).

These are the start dates for each cycle

- Cycle 4 - 13 Nov 2017
- Cycle 5 - 5 Feb 2018
- Cycle 6 - 11 Jun 2018
- Cycle 7 - 10 Sep 2018

Using this information I plotted the following graph looking at the time at which people left and the week number. I expect that as the weeks in a cycle go on we should see distinct groups week numbers leaving. The Y axis is the hour someone left, This was done to help avoid overlaps of points (dosnt add anything to the graph)



When adding a line of each of the course start dates we see that there seems to be alot of people dropping out at the start of each cycle. We also see that that in cycle 6 and 7 there where a few people that dropped out before the course began. There wasn't any distinct cluster based on week. What we see is that people don't work on the course at the same pace, there might be some people who rush through this and some that takes time.

What is odd is that in cycle 6 is that there is a person who has apparently completed week 3 before the course even began.

Whats going on with the person who left at week 3 before the course even started In cycle 6 we see someone left at week 3 before the course even started. How is this possible?

We see that there was 7, 11 people who left the course before it had stared. All these people didn't complete a week par the one person that finished week 3.

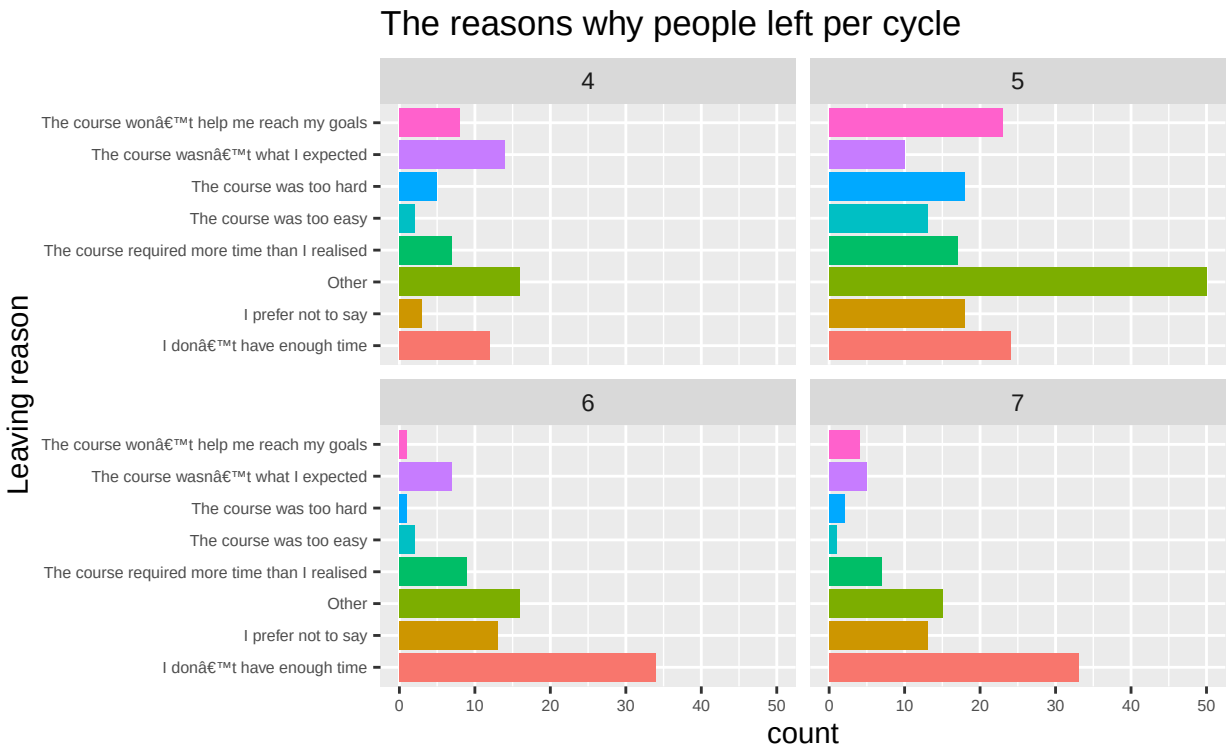
We see this person is user 353eaf1b-cff8-4d9c-98d4-b9a736c17e73 who left the course on the 2018-05-09 03:54:35 however there last step is was done on the 2018-07-15 16:53:37 which is 2 months after they left. This should be possible. This user only has one record in this data so an error has occurred in the recording of this data.

This has lead me to thinking that there may be more error like this so I will need to do a check for more of these "impossible" leavers

Assumption 3

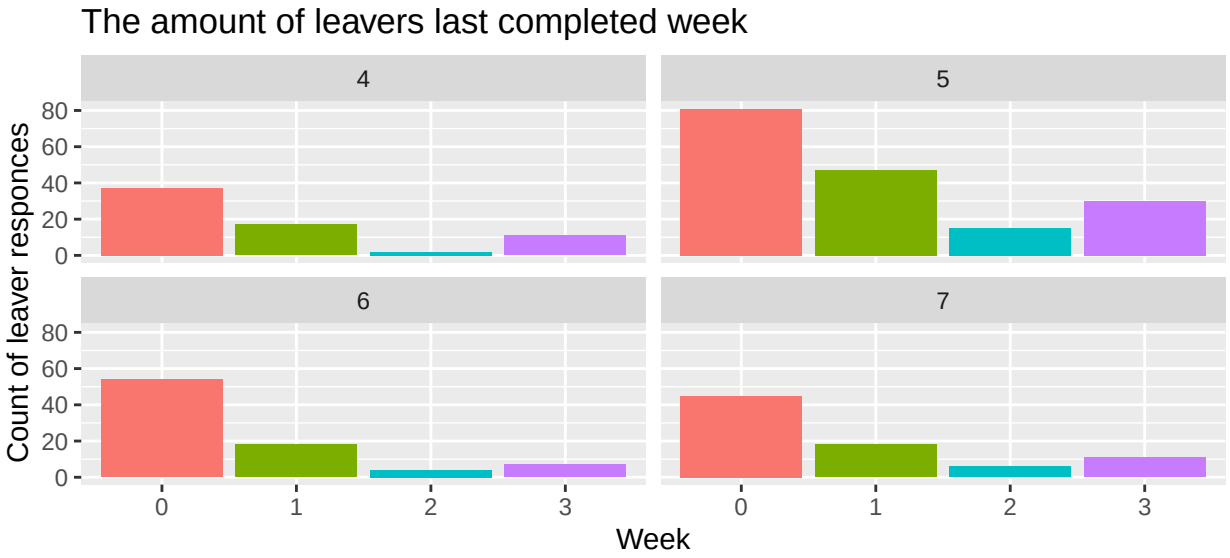
In the whole dataset there seems to be 9 users who left the course but still where able to complete a set after they left. As this is 8 user spread over the 4 cycles I will assume that these are errors and remove them in order to keep the data tidy and bit less error free.

Reason for leaving per cycle. With the data cleaned a bit more I decided that I will look into the. The reason why people left each cycle. As we saw previously alot of the leavers who left a response happened in cycle 5. I wanted to figure out if there was a reason for this or if this cycle followed the trends of the other ones.



If we split the results into the cycle they where achieved from we see that in cycle 6 and 7 the majority of people left because they didn't have time. We also see that in cycle 5 the dominate reason was for some unknown (other). This is the cycle that had the most drop outs so this suggest that this cycle contained a 'unique' issues that might of only occurred during this cycle and cause alot of people to leave.

Weeks leaving per cycle. Next I wanted to break down how many people left in each week per cycle. I expect that there be alot of people leaving at the start of the course but that it would decrease as the weeks went on. By week I mean the teaching week and not the weeks after the course starts



Looking at the graph we see that for all cycles, a large majority of people left without starting the course. We also see that all cycle follow the same trend 1st week has the most leavers, 3rd week and then the second

week. There isn't alot of people that left in the second week suggesting that this week keeps its users quite engaged.

Reason for leaving per week. As we see that throughout each cycle the amount of leavers per week is roughly the same I wanted to see what the primary reasons why people left each week.

When turning the amount of responses into a graph of percentage we see the following.

For weeks 0,1 and 2 the main leaving reason was that the users didn't have time which makes sense. They are likely to leave the course early if they know they don't have time than try to continue.

In week one we see that more people said the course was too difficult when compared to those who said it was to easy. We also see that in week 1 and 3 both a roughly similar suggesting the amount if challenge this week. In week 2 we see that there are no people leaving because the course is difficult suggesting that the difficulty of this week should be improved in order to keep people from leaving.

We see that the highest percentage "the course wasn't what I expected" occurs in cycle 1 which is what you would expected as people would join the course do a bit for the first week and that then decide that the course isnt what they expect.

We see that an overwhelming majority of people left in week 3 said the reason was other, this suggest there might be something that has happened in week 3 that caused people to leave.

Time between leaving While most of the people who leave the course don't complete a week there is still 174 people who did leave after at least starting the course.

I wanted to find what the time between the last step that a user completes and when they officially leave the course. I felt that this could be useful as if we know how long it take a user to leave the course after there last step ,something could be done to prevent it.

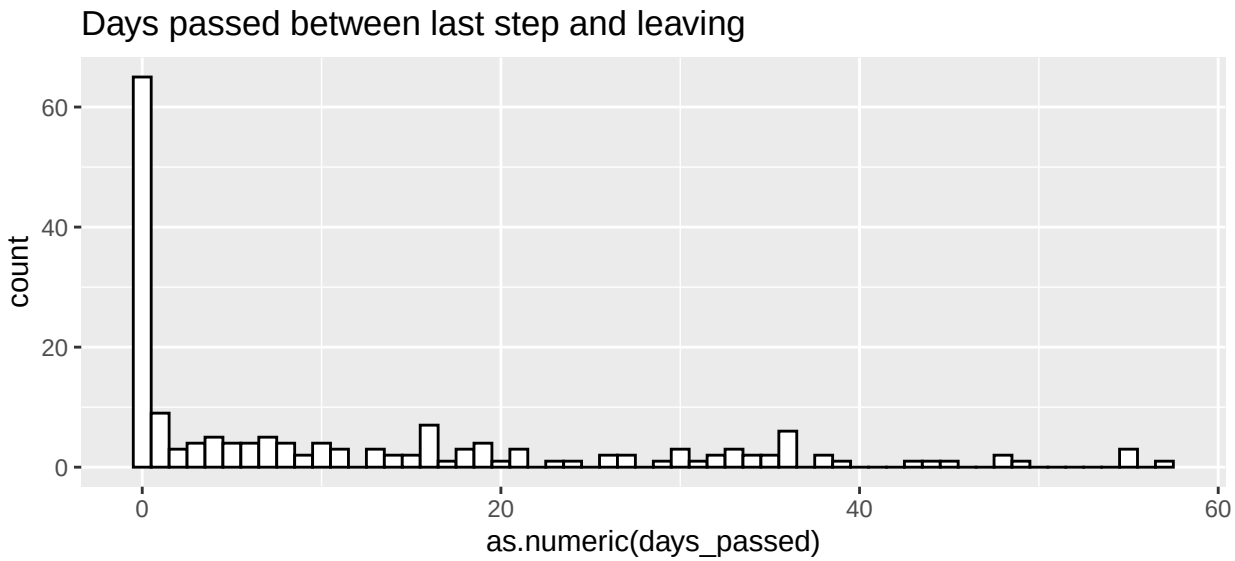
For this a new sub table was created that works out the time in minuets between the users last step and the users leaving data. Also for simplicity I grouped the minuets into groups of how many days had passed between the two events.

When working out the times between the two events but rounded down to number days that passed we see the following.

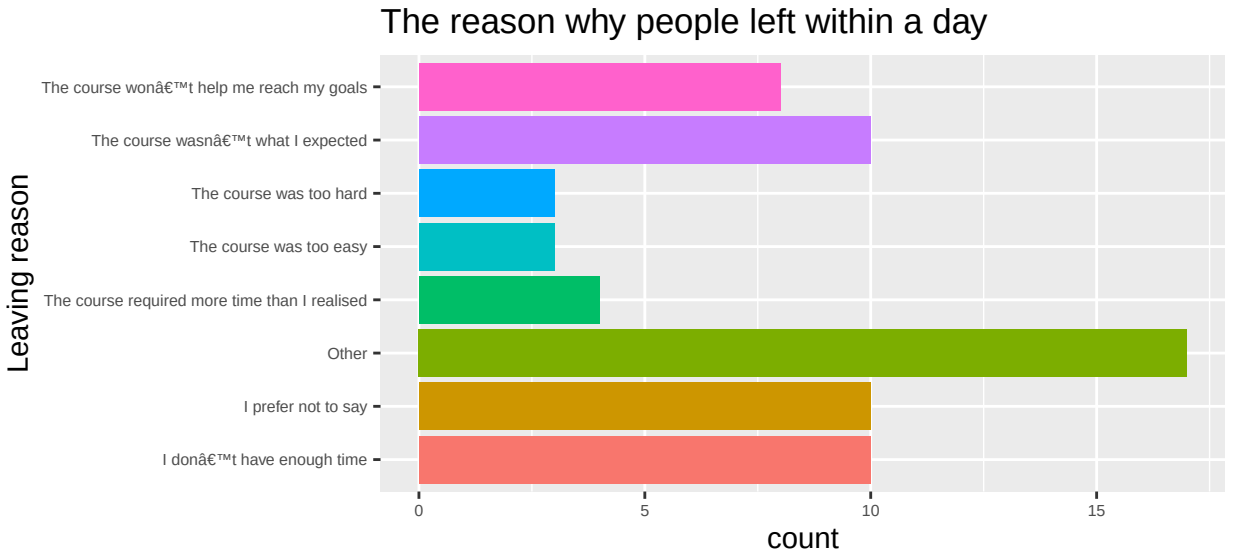
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    5.00   11.99   19.00   57.00
```

We see the median number of days that had passed was 5. This means that that half the people leave the course within one working week of completing there last step. What is alarming is that 25% of the people leave within the 24 hours of completing their last cycle step. Suggesting that most of the time the user dosn't need much time to think about leaving they finish the last step and then idmedantly think about leaving.

Visually representing this as a histogram show us the following.



We can see that most of the the people leave within 24 hours for them completing there last step. This backs up the idead that when a user wanst to leave they are most likely do this after finishing a step. Looking a the reasons why. we see



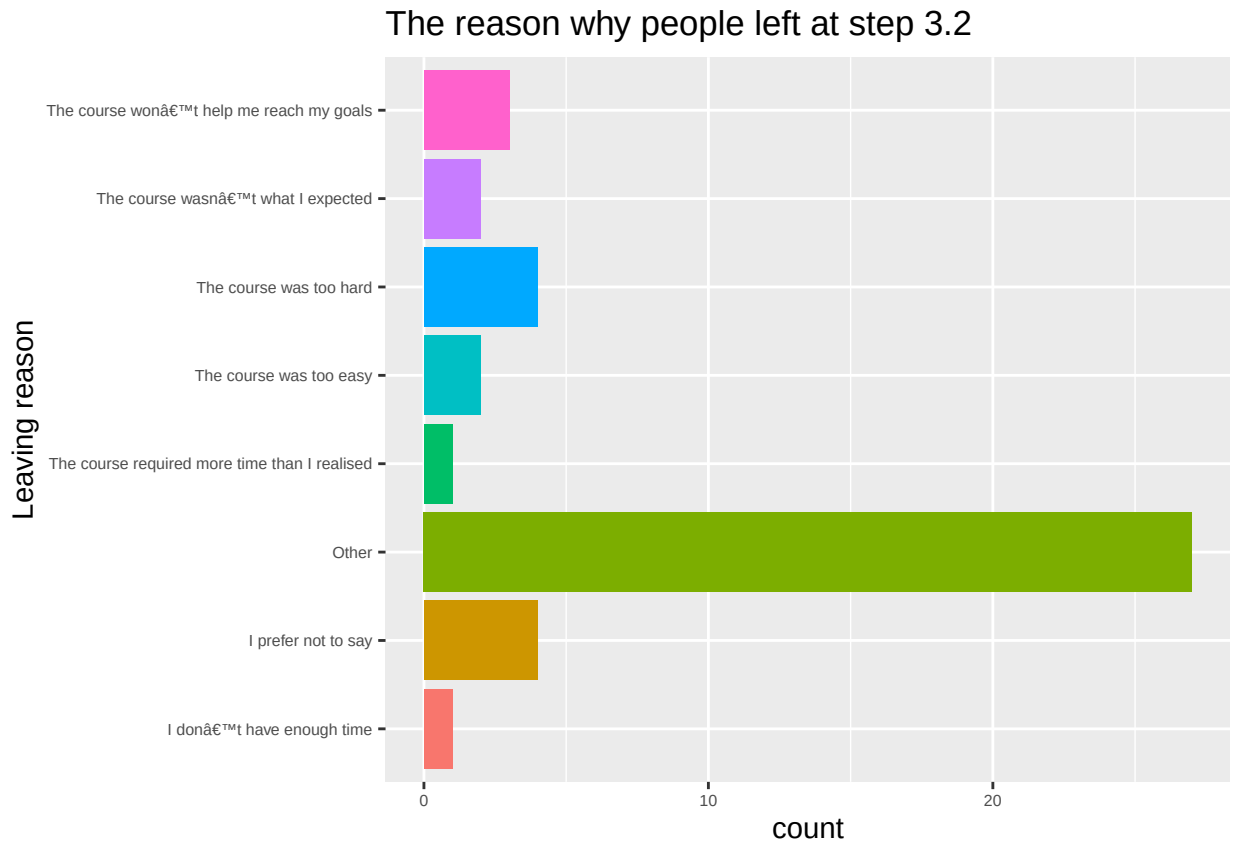
There is alot of people who left who didn't give a spec reason (other / I prefer not to say). We see that there is a few people That it has anything to do with difficulty. We see that there is a high amount of people that left because that the felt that the course didn't help them or that it wouldn't help them achieve there goal. I somewhat expected this to be high as if a person didn't think that the course would help them you wouldn't expect them to stay around for long.

Why is there an increase in people leaving in week 3

I really wanted to see why people where leaving in the 3rd week. When we did a count of the number of people who left at each step we saw that there was alot of people that left within the first few steps but there also seemed to be alot of people leaving at step 3.2. This happened in all cycles.

Looking at the courses step 3.2 is a video (Devices in the future home VIDEO (03:26)).

However what should also be mentioned is that the final step is 3.20, so there is a possibility that there has been some error in recording the data (people who finished the course are accidentally put down as leaving).



Looking at the reason given we see that the reason is overwhelimily "other". This suggest that there is a really unique issue with this video that is causing people to leave. Maybe its offending people?

Evaluation of cycle one

What I found and what could be done

Whilist I am happy that my analysis is on a cleaned dataset there where a few assumptions that had to be made.

After cleaning up my data set and getting a grips with it I decide that I wanted to Evaluate what I had found and how it could be helpful in answering my business question into "give useful insights into why people decide to leave the course".

- I have broken down alot of the timings when people leave and the most common given reason. For the most part it is either that the user dosn't have time or that they don't want to say. This was somewhat guessable.

- People who left during the first week had a highest rate of people saying the course was not what they expected suggesting that there may been to be a better description of the course so people know that there getting into
- We see that there is in nearly all situation where we look at the response The amount of people saying the course was too easy or too hard was roughly the same suggesting that the difficulty isn't too much of an issue. However week 2 might be a bit to easy so the difficulty may need to be increased in order to keep users engaged.
- The majority (around 52%) of the people that leave don't even complete a week, suggesting that people enrole on the course but don't do anything.
- If a person is going to leave they are likely to do it within a day of finishing within the day.
- For some unknown reason alot of people leave at step 3.2 which suggests that the video may be offending people / causing them to leave.

If i where to Rank these I have suggest that step 3.2 is looked into. If people are already at week 3 we would expect that they would be wanting to complete the course. If step 3.2 is turning people away then something should be done to change that.

Review process

As I wanted to find out why people left I ended up creating some similar looking graphs.

I felt like all my analysis was necessary but there where a few things that I couldn't include in this report due to space. I should of cleaned then explored and not both at the same time as where a few time where i had to go back and update things.

I found a fair bit that I could use to answer my business question in regards to why people leave at but only at what times. However we don't really know who the people leaving are. If we could get information about who to was leaving and why to see if there was any common grounds. Then more measures could be put in place to keep user from leaving the course. This gave me the idea to do another cycle to try and look into this.

Data Understanding, Cycle two

As I want to see who these people are that left these reviews I decided to look through all the data again to see what I could work with. I felt that the enrollments data was the perfect data set for this. As we have enrollment data for all cycles I decided to merge the enrollment data for cycles 4 to 7 together then I then merged a copy of that dataset with my cleaned leave reason database.

Explore Data, Cycle two

Doing this merger gave me the following data to look at Gender, Enrolled time , role, age range, highest_education_level, employment_statue, employment_area and detected_country.

About the users - who are they

Looking at the columns gender, Country and age range we see that the majority of these columns are dominated by Unknown data. This may be caused by the users not wanting to enter any personal data about themselves or some issue when collecting the data.

Gender

```
##
##  female    male   other Unknown
##      63      58       1     281
```

Country

```
##
##      AO      AU      BG      BR      CH      CR      DE      EG      ES      GB
##       1       4       4       1       1       1       4       1       4      55
##      GH      GR      IE      IN      IR      IT      KE      MM      MX      MY
##       1       1       1       2       2       3       1       1       1       1
##      NG      NL      NZ      PE      PL      PS      RO      RU      SA      SY
##       1       3       3       4       1       1       1       1       2       1
##      TR      UA Unknown      US      WF
##       1       4     279      10       1
```

Age range

```
##
##     <18     >65   18-25   26-35   36-45   46-55   56-65 Unknown
##       2      16      10      19      18      31      20     287
```

We can see that not all the Unknowns are the same size so this tells us that it might not be an error in collecting data as you would expect that all data the same rows would be unknown and not different sizes. This suggest that there is a fair few users that are not wanting to give information.

##Assumption 4 When given an option too, the users are choosing to not give information. So the NA's aren't caused by an error in processing the data. Its just the users not wanting to give up information

Dected country compared to users entered country

There is two columns that both contain country data, country which I'm guessing what the user entered and detected_country which is probably and automatic check that has been run without the user knowing and recorded. When looking at detected_country there is alot more of a variety in the data detected_country than country (as there inst any unknown values). Doing a check of all known counties vs the results of detected_country we see that there is 108 similarities and 8 Differences suggesting that detected_country is more reliable than country data.

About the users 2 - education and employment

After looking at the area of highest_education_level, employment_status and employment_area

highest_education_level

```
##
##       apprenticeship  less_than_secondary         professional
##                    1                    2                   14
##            secondary             tertiary    university_degree
##                   21                    8                   39
## university_doctorate   university_masters              Unknown
##                    1                   34                  283
```

employment_status

```
##
## full_time_student  looking_for_work       not_working           retired
##                 7                 9                13                22
##     self_employed        unemployed           Unknown working_full_time
##                12                 6               283                39
## working_part_time
##                12
```

employment_area

```
##
##     accountancy_banking_and_finance armed_forces_and_emergency_services
##                                   2                                   2
##  business_consulting_and_management        charities_and_voluntary_work
##                                   4                                   5
##           creative_arts_and_culture               energy_and_utilities
##                                   7                                   2
##       engineering_and_manufacturing             health_and_social_care
##                                   3                                   4
##       hospitality_tourism_and_sport         it_and_information_services
##                                   1                                  15
##                                 law        marketing_advertising_and_pr
##                                   3                                   4
##               media_and_publishing                       public_sector
##                                   3                                  12
##                   retail_and_sales         science_and_pharmaceuticals
##                                   1                                   1
##              teaching_and_education             transport_and_logistics
##                                  12                                   1
##                             Unknown
##                                 321
```

Looking at these variables we see that there is a large amount of unknowns in each columns. I think this is primarily due to users don't wanting to say what they do, maybe there is inst the option to exactly state what the user does/is.

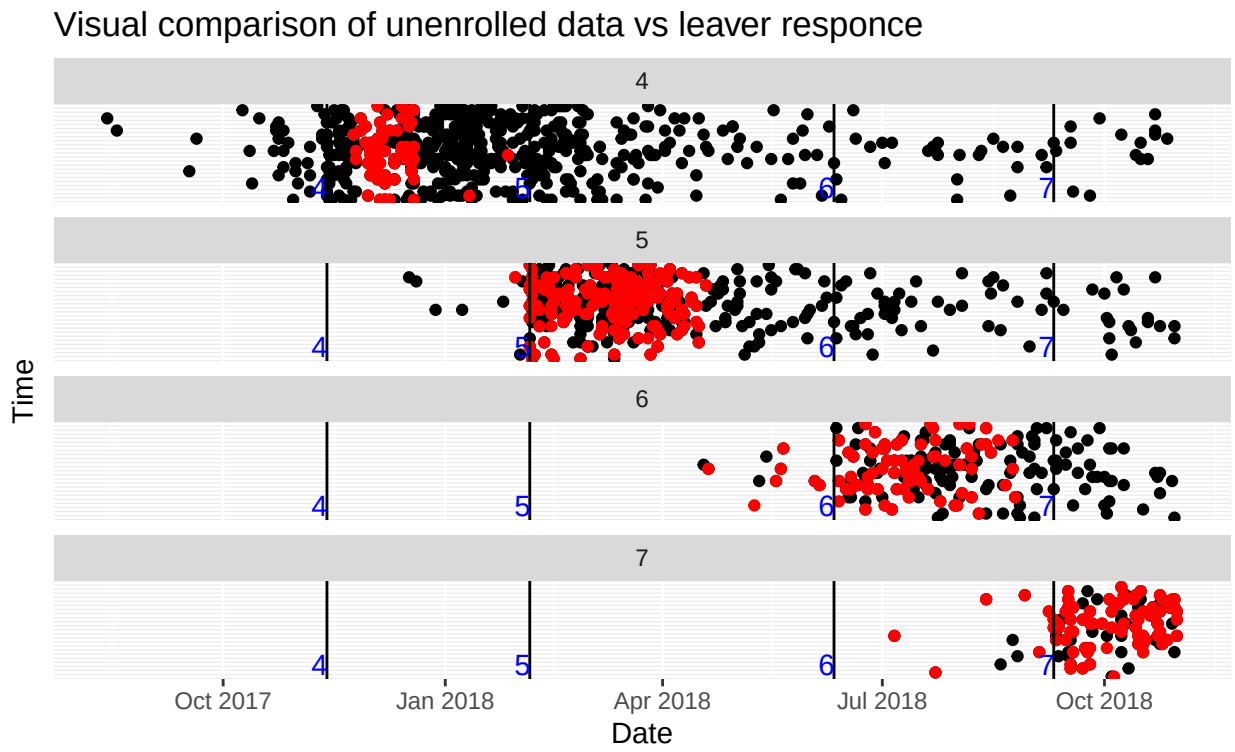
The unkown problem The amounts of unknown in the user data dwarfs any other variable in that data, especially when there is alot of options that a user could be (for example employment_area).

Two leaving dates The merger gave me two users leaving dates. The Unrolled time from the enrollment data and the leaving time from the leaver response data. I wanted to check the if the amount of people that gave a response to leaving and the number of people unrolled from the course was the same number. Looking at the combined user data we see that there where 1279 users who left the course which is alot more than those who gave a leaving response. However the total size of the combined enrollment data was 13053 meaning we have 11774 rows with no unenrollment data.

```
table(Unenrolled_combined$Cycle)
```

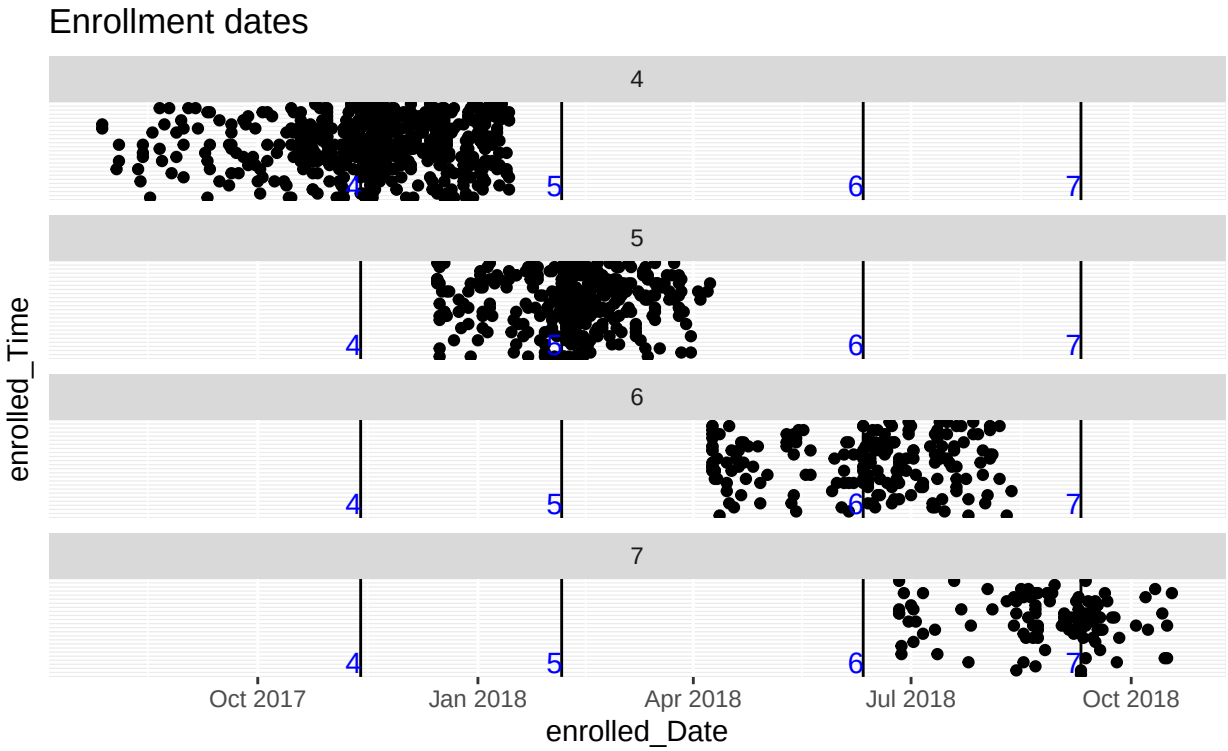
```
##
##   4   5   6   7
## 560 407 203 109
```

In previous analysis we see that the most the user leaver reports where in the 5th cycle. This would suggest that most of the people that left would be in cycle 5. However this isn't the case as we can see the majority of people with unenrolment dates are in the 4th cycle.



I wanted to visually compared the two set of leaving data, the enrolled data with an unenrolled_at date and the leavers response data set to see what they would look like against one another. We find something interesting with by doing this. What we see is that there as the cycles go on there is alot less people leaving so the difference between the amount of unrolled_at and leavers response becomes smaller. What we also see is that alot of cycles have people unrolling surprisingly late after the course (for example there are people in the 4th cycle that unenrolled in the 7th cycle)

As we saw with the Visual comparison of unrolled data vs leaver response along a time line we see that there is a lot of people unenrolling months after there course had ended. I thought (though unlikely) that this may be an error in the data being recorded (i.e people are being recorded in one cycles database when they actually attended another cycle). I decided to check when people enrolled for a course. If people had enrolled way after that cycle ended then there had obviously had been some issues.



Looking at the time scale of enrollment dates we see this doesn't hold, there is no instance where a person enroles for a course after it had finished. What is weird is that there does seem to be some overlap in enrollment times. For example there are people who enroll just before cycle 4 ended and at the same time there are people who are enrolling for cycle 5. This may mean that a user can apply for course whenever they want to and then have as long as they want to complete it.

Assumtion 5

As we see we have people that have an people with an unenrollment date and people who haven't. Based on this I am going to assume the following. People without any unenrollment date are still working on the material or have just left without formally leaving (they just logged off one day and haven't come back). This means that I will assume that the unenrollment date is for both those who have left the course and those who have completed the course.

Data Exploration, Cycle 2

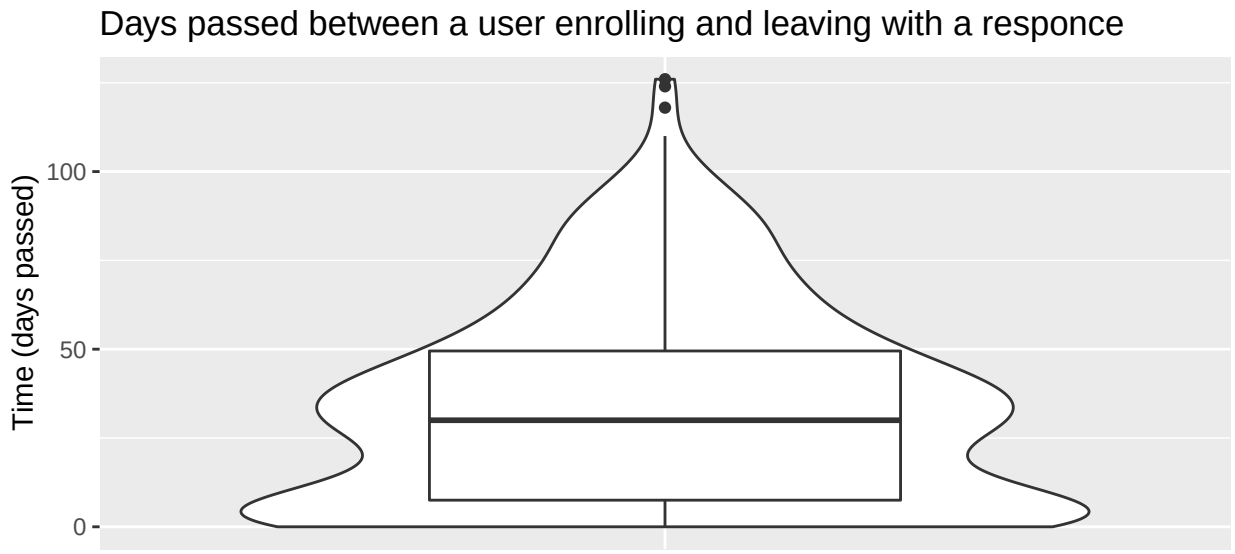
As we have seen alot of the data that I wanted to look at that help define who a user was mostly made up by NA values. This meant that we only had a little bit of data that we could actually use if we wanted to look at things like how age effected why people left. I did do checks on these graphs but it didn't yield any decent results as it wouldn't make sense to state how an over all group of people found the course based off what one or two people said.

Time between enrolling and leaving

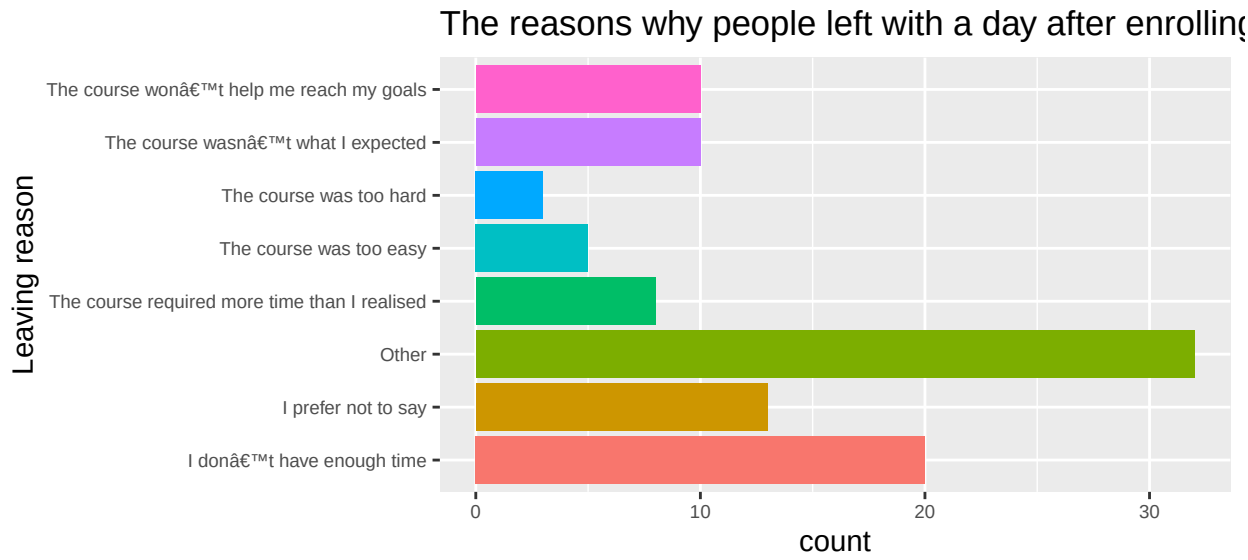
Whilst not exactly the main reason I merged the enrollment and leaving data together, it did give me the oppurinty to find out the time between when a user joined the course and then left the course and gave a

response. Finding out the typical time for someone to leave the course could be interesting as it means that measure could be put in place to try and stop this.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.50   30.00   33.07   49.50  126.00
```



As we can alarmingly see, from the violin and the box plot, a lot of the people within a week after enrolling the.



When looking at the reason why we see that there is alot of people who say that they left for the "other" reasons, As expected of people who would leave just as a course began not having enough time is another big chunk of people leaving the course.

Evaluation of secound cycle.

Due to the poor quality of the results captured in the enrollment data base there wasn't too much that we would get from this set of data. In all of the columns that i wanted to look at (age, role, education) Na

values took up a majority of all the total results, splits the remaining results across around 5 ish different variables we would only have 1 or 2 leaver response in each category which i didn't feel like it was smart to categories a full group from people by what one person said. There is not too much I could do to fix this without doing further analysis using extra data. I just had to leave this data as unkown.

What I found

Though i didn't find out what I was expecting I was able to use the enrollment time to good effect.

- In the enrollment data there is 1279 users with an unenrollment date meaning that there either allot of people that are unenrolling from the course and decide to leave the course. Unerollment date could also be a combination of people who finished the course plus left the course meaning. If this is true that means in cycles 4 to 7 there is 11774 users who didn't complete or officially leave the course.
- When looking at how long it takes a user to officially leave after they joined the course it takes on average a month. However we also see that 25% of these people leave with a week, with the primary reason being other.

Whilst not what I was looking for I do believe that both these meet the business object "give useful insights into why people decide to leave the course" as the first point hints that either that the we have a bunch of that leave but don't give a response or we have alot of people that are starting the course but dropping out with out officially leaving. The other point suggest that when a person enrolls there is a fair bit of time before they un-enroll so the creators of the course may have more time than previously thought to get users more engaged in the course.

As I decide that this would be my last cycle and that I would move onto Deployment

Deployment

Conclusions for the business

Whilst I did uncover a few bits of information that would give us insight into why people leave the courses the main ones that I want to highlight is that

- Alot of people are leaving at step 3.2. This is a video and the main reason people give when leaving is other. This suggest that something is wrong with the video and that it may be offending/upsetting people to the point they want to leave. This should be looked to keep wanting to finish the course.
- There is alot of people who leave without starting the course. This should be immedantly looked into as if you can get people excited for starting the course Then they will likely give the 1st step a shot and that may get them hooked.
- Alot of people are likely to leave just after they finished there last step, whilist this may be difficult to stop its possiable that mabye sending a message after a user completed a step saying whats comming next and that there doing well may keep them invested.

Review of business objectives

I feel like my project was able to meet the business objective of "give useful insights into why people decide to leave the course" as I was able to find several insights into why people leave the course. However this pirmaly have to do with the reason why people leave given a certain / point in time and nothing else. More work could and should be done to look into who is leaving or if there is any other causes that make people leave the course. We see that most of the time that a user says the reson they left was other, it would be nice to know what that that other refers to.

Review project

Whilst I felt that my first cycle went quite well as I was able to clean up the data and was able to find find alot to do with "why people decide to leave the course" as the leaver response data contained alot of useful facts and figures that could help us. Unfortantly I ran into some issues using the enrollment data, there was too much missing data for me to make any real use out of it apart from doing some stuff in the enrolment data which is a same as finding out who excatly is leaving could help us predict why a similar user may leave in the future.

Summary of results

I felt like that my analysis was good, I didnt have to make that many assumptions and those assumptions that I did make I gave evidence to why I think my assumption would be correct. Everything I made is storted within the project so all my results are reproducable meaning that if someone whats to look over my code they can

Future data mining

If we could get more information about who is leaving we could find key insights into why different groups of people and update the course accordingly.