Executive summary

Callum Simpson

B6030326

The Crisp DM

I feel like my processes for the crisp dm model worked well.

After creating my question for what I was going to look into (why people left the course), I went over all the data to gain an understanding of it. The data understanding step was helpful as it made me thoroughly go through all the data meaning that from the start, I had a good idea of what I was working with and how I could use it.

Once I had selected the data that I wanted to work on I went through all of the different columns to see if there were any NA values or any issue with the data. Like seeing if there was any duplicate IDs in the same cycle. Once I had counts of rows, I went about cleaning my data which was fairly easy to do. From there I went about using my data to create exploratory graphs that help me answer my business question. If I felt like I had enough to answer the business question I went to the evaluation step to review. On my fist cycle I had an idea of a useful second cycle that I could do in that could help me better answer my question. Do this was simple as I didn't really need to preplan anything. I could just incorporate it after my current cycle.

I liked how the evaluation step works. How it allows you to compare your results to your initial business task and decide whether you are done of if you want to do a new cycle to try and uncover some new information/ try and solve your question.

I also liked each step somewhat allows you to review what you had done and why you had done it. It made me continuously think about my work and that help me improve in any analysis. For example, in the Data preparation step I would prepare my data to make a graph and then the results of the graph would make me think if I wanted to go back (which would be easy to do) and prepare my data again to try and create a different model.

I also liked how each section followed into one another as it meant that using the Crisp-dm was easy and frustration free.

An issue I had with the CRISP-DM model was the modelling Phase. This primarily due to modelling being for doing machine learning algorithms. This project was more of an exploratory task so we couldn't really utilize this step fully. I instead used this step to create graphs.

I did have difficulty with the Deployment step as we were really "deploying" in this project in the same way you would do in a crisp DM model. Instead, I just wrote about key finding and what the business should do.

The Data

The data I used was the enrollment data and the leavers response data. Whilst it's a shame I didn't have leavers responses for all cycles (so I could have more to work with) I was still able to do a lot with merged data set. It didn't have many issues with NA values as I was able to make an assumption that help me understand what the NAs meant (week zero). I cleaned up this data so to a point where I felt the I was safe using it as I had gotten rid of most of the errors. The only issue I had with is it is that other is a very generic response that could mean anything. If the user could tell us what was wrong then I feel like I could do some more detailed analysis.

The number of NAs in the “user data” enrollment data mean that it was really difficult to work with as I couldn’t really use it to get a good reason why groups of people where likely to as I only had a handful of results per group. If it was made so that a user had to give that information Then I feel like I could some really analysis on who was leaving and the reason why which would give us Cool insight into why people left.

Data assumption

During the course I had to make a few assumption

- 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. This also suggest that courses where updated in each cycle
- The most common reason for people leaving cycle 5 is other suggesting that something unique happened in step 5
- Most of the NA values in the leaving responses are actually “week zero” users. So people who left the course without completing a step. Because of this I set these values to be week 0 and used them as such when looking out leaving reasons and week
- The reason for people leaving at step 3.2 is because of the video. This suggest that there is something with the video that is causing people to leave.
- People who don’t have unenrollment date haven't left the course. I assumed that these people are “unofficial leavers”. People that leave the course by walking away and not leaving the response.
- All the NA data in the “user data” columns in the enrollment error is caused by the user not wanting to give a answer. Meaning trying to guess who the user is would be extremely difficult unless they told you.

GGPlots

I felt that my ggplots went well, I feel like all my graphs work well and clearly descried what is going on. I was able to do a wide number of different graphs so I have a Varity of different plots to look at.

Project Template

For the first time using project template I think it really well. I have stored all my reports in the reports section, data in the data section, src contains a bunch of source code for my reports and munge contains code relating to database creation. I found it useful as it allowed me to ensure repeatability whilst also keeping in my database tidy.

GIt

I found git extremely useful as if I had a mistake I could go back and get the previous work code.