<u>Graham Cole – a3030758 – CSC8631: Exploratory Data Analysis</u>

<u>Future Learn Project – Design and Implementation Report</u>

<u>Executive Summary</u>

The project required an analysis of data pertaining to 7 unique runs of an online course provided by Newcastle University on the Future Learn platform. My analysis focussed on two lines of enquiry; creation of a demographic profile for each archetype class and an interrogation of course progress, again through the lense of an archetypal categorisation. My findings showed the following:

1. Demographic Profile.

   Despite incomplete information across the data set, my initial analysis showed that archetypes could be grouped by demographic information (age, gender, employment status). This was most notable for the Vitaliser archetype which I found to be predominantly female, self-employed and in the age ranges 46-55 and >65. Further investigation may yield similar stratifications, perhaps driving marketing activity or tailoring of course content.

2. Course Progression

   Whilst the archetype classification did yield some insight (Preparers and Hobbyist performed significantly better than their counterparts in 2017) the analysis showed a significant deterioration in the performance of all archetypes in 2018 as compared to 2017. This may warrant further investigation into course design, delivery and/or the duration of participation across the learner population.

<u>Crisp-DM Methodology – *Help or Hindrance*</u>

Stage 1 – Business and Data Understanding

I very quickly identified the archetype classification field as being a potential focus for my analysis. Despite the limited number of learners that had this attribute, I felt that it could provide a good focal point, potentially enabling an appreciation of the characteristics attributed to each archetype matched against a summary of their performance.

I was able to discuss this with Sam Flowers (the “business”), requesting further information about the approach taken in implementing the archetypal classification. Based on his positive feedback, I decided to begin the analysis.

On reflection, I would now limit my analysis to one of the two lines of enquiry as I feel that, whilst my results did show early signs of trends, they were both too shallow to properly explore any signals in the data. I would also seek to engage more frequently with the business which may have helped me to narrow my focus.

Stage 2 – Data Preparation

In order to analyse the data, I first needed to consolidate the separate data files into a master data table which included all information pertinent to my enquiries. I immediately discounted runs 1 and 2, as they did not contain any information regarding the archetype classification. Having done this, I consolidated similar files from each run into a master file per data type (i.e. consolidating all enrolment data files from runs 3:7). I was then able to join this data with the archetype data, removing those learners which did not have an archetype value. I conducted this process twice; once for the demographic data and again for the course progress data.

The data wrangling phase of this project was particularly challenging for me as it required use of several packages in R that I was not familiar with, mainly dplyr, ggplot and lubridate. However, I am pleased with the progress that I made in this regard and feel more confident in utilising the broader functionality of R Studio. One point of improvement could have been to remove those variables which were not used in my analysis to further streamline the data set and aid reproducibility and legibility.

Steps 3 & 4 – Modelling and Evaluation

Due to the nature of the variables which I chose to analyse (categorical and discrete numerical) I did not need to construct a model in the traditional sense. Instead the Modelling stage of my analysis focussed on the collation and scrutiny of variables in line with my overall research aims. Using the combined data set described above, I was able to group, sort and filter the data to investigate trends within the archetypal classes. Building on these groupings I utilised simple statistical measures of central tendency and spread to provide insight into course progress.

I adopted an agile approach to my evaluation, basing the proceeding evaluation on the previous results. For example, in grouping the learners by archetype and age range and plotting this graphically, I identified significant groupings of the Vitaliser archetype within the 46-55 age range. This led me to investigate this cohort further to provide clarity and comfort that these results were indeed representative of the data.

Having seen my counterparts' presentations in our shared VIVA session, I think that there was more that I could have done to apply more sophisticated statistical analysis, in line with limiting my analysis to one of the fields of enquiry I identified at the beginning of the project. There could, for example, have been an opportunity to plot two variables against each other (for example age and gender) to identify correlation.

Step 5 – Deployment

I chose to deploy my findings through an HTML RMarkdown document and present this document in my VIVA session. Again, having seen colleagues' deployment on Shiny, I believe that my analysis could have been better communicated via this more interactive medium. That said, I believe that my use of colour coding the bar graphs to show relative frequency by archetype within the variable analysed was effective.

In addition, I did not have a strong grasp of the figures that I was presenting in my VIVA due to only finishing my analysis on the morning of the presentation. Had I limited my lines of enquiry from the outset, I would have had more time to better understand my findings and deliver a more informative message to the business.

<u>Conclusion</u>

Overall, I feel that I met the project objectives reasonably enough. As set out above, limiting the scope of my analysis may have yielded more time to allow for a more in depth and meaningful look into one area. My continuing struggle with R Studio meant that a disproportionate amount of time was spent collating and scrutinising the data however I feel that I have improved significantly in this regard over the course of the project. I also had trouble enabling Git Hub on my computer which meant that my version control history was not representative of all of the stages of my work.