CSC8631 Project

Luke Battle

22/11/2020

Introduction/Business Understanding

In this report, we will use the CRISP-DM method to analyse a data set. This will involve us running through several key steps; *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evalutaion* and *Deployment*.

Business Understanding

The first step in performing our CRISP-DM analysis is to gain an understanding of the business motivation for the analysis, with the primary goal of this analysis being to enhance Newcastle University's online resources. Specifcally, for the past several years, Newcastle University has ran a free online course called *Cyber Security: Safety at Home, Online, in life*. Several data sets have been compiled each time the course has been ran, detailing several facets of the user experience throughout the course. Newcastle University seeks to utilise this data to enhance the quality of material taught outside of the classroom, and ultimately promote learner engagement with the course and Newcastle's wider online resources. To achieve this, different elements of the collected data will be examined over consecutive runs, with the aim of identifying any trends both within and between them that may indicate an area where user engagement could be improved.

Data Understanding

Next, before beginning our analysis it is important to fully assess the data, so that we can make informed decisions about the portions of the data that should constitute the foundations of our analysis. In total, Newcastle University has provided 53 csv files that describe a range of features of user interaction with the MOOC over 7 runs of the course. These runs span from September 2016 to September 2018. For every run there is a file detailing "Archetype-survey-reponses", "Enrolments", "Leaving Survey Response", "Question Response", "Step Activity" and "Weekly Sentiment Survey Response". In addition, from run 2 to 7 a file on the "Team Members" is included, and from run 3 to 7 there is also a file on the "Video Stats". However, the data quality within some of these files is quite poor. For instance, "Leaving Survey Response" (**NOTE THAT THIS HAS BEEN USED IN LATER RUNS**), "Archetype Survey Response" and "Weekly Sentiment Survey Response" are blank except for their column headers, so they will not feature in this analysis. The "enrolments" data set does contain data, however the quality of it is poor. For 88.02% of the individuals in the data set, the only data available is the detected country of origin and enrollment date. These individuals are missing useful demographic indicators such as "gender, age range, education level and employment status. Due to the poor quality of the data this will not be used in the analysis. Out of the remaining data sets, video stats and question response may be a good place to start in assessing the efficacy of the MOOC course.

The video stats files are included for course runs 3 to 7. The files each have dimensions:

```
## [1] 13 28
```

With the following variables:

```
##  [1] "step_position"                  "title"
##  [3] "video_duration"                 "total_views"
##  [5] "total_downloads"                "total_caption_views"
##  [7] "total_transcript_views"         "viewed_hd"
##  [9] "viewed_five_percent"            "viewed_ten_percent"
## [11] "viewed_twentyfive_percent"      "viewed_fifty_percent"
## [13] "viewed_seventyfive_percent"     "viewed_ninetyfive_percent"
## [15] "viewed_onehundred_percent"      "console_device_percentage"
## [17] "desktop_device_percentage"      "mobile_device_percentage"
## [19] "tv_device_percentage"           "tablet_device_percentage"
## [21] "unknown_device_percentage"      "europe_views_percentage"
## [23] "oceania_views_percentage"       "asia_views_percentage"
## [25] "north_america_views_percentage" "south_america_views_percentage"
## [27] "africa_views_percentage"        "antarctica_views_percentage"
## [29] "Run"
```

Where each individual in the file is given by the video that the variables relate to. Which are the following:

```
##  [1] "Welcome to the course"
##  [2] "Why would anyone want your data?"
##  [3] "Preserving privacy in cloud storage: privacy by design"
##  [4] "Staying safe online: personal perspectives"
##  [5] "Privacy online and offline"
##  [6] "Welcome to Week 2: payment security"
##  [7] "Exploring vulnerabilities in online payments"
##  [8] "The million dollar contactless payment"
##  [9] "The evolving arms race of payment security"
## [10] "Welcome to Week 3: security in the future home"
## [11] "Exploring security: biometric authentication"
## [12] "Exploring security: the Access Control Live Lab"
## [13] "Devices in the future home"
```

For the question response data set, we can compute the same high level analysis. However, unlike the video stats files this file has been prepared for all 7 runs of the course, and is not in exactly the same format. Each file has 10 columns, however the rows in each data file is dependent on how many learners took part in the corresponding run, and so varies between files. In total, all 7 files have 176,463 rows.

The data sets used in this report are details of user engagement with the videos and quizzes that feature within the course. Within the video data, I have made the assumption that all videos are the same in each run. This assumption has been made on the basis that the video duration is the same for every run of the course, thus it is unlikely that any changes have been made.

Both of these elements will be explored and examined both individually and in relation to each other, with the ultimate aim of providing feedback to Newcastle University on how it can enhance the user experience of the course. This analysis will be constructed using CRISP-DM methodology and, as such, will be structured to reflect this.

Data Preparation

In order to prepare the data for analysis we will first seek to compile all runs into one file

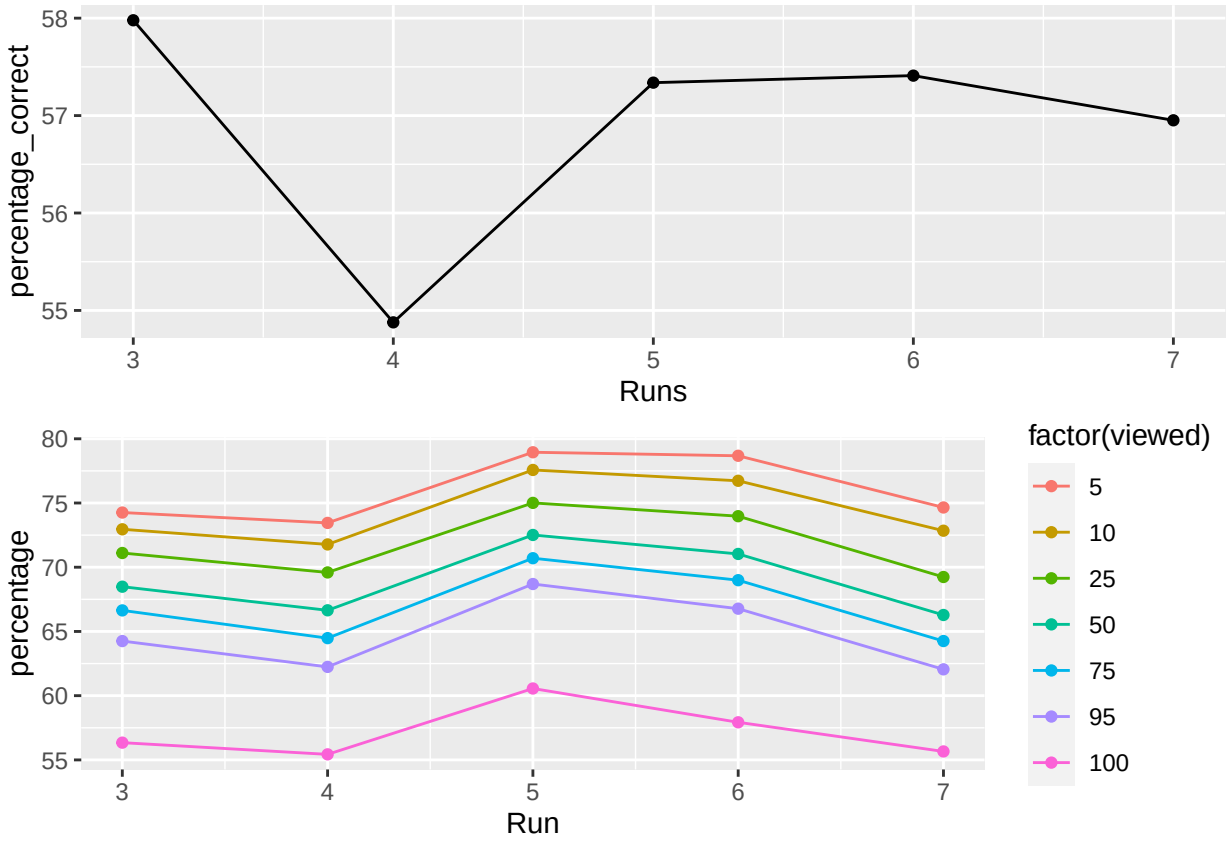
Modelling

Now that both data sets are appropriately formatted, we will begin our analysis of the data. First of all, we will investigate if there is a relationship between the percentage of people who watched the videos in the course, and their results in the multiple choice quiz. To examine this, we will calculate the percentage of questions answered correctly for the each run of the course.

```
knitr::kable(correct_questions, col.names=c("Run", "% Correct"), align = "c", digits = 2)
```

Run	% Correct
1	54.78
2	58.67
3	57.98
4	54.88
5	57.34
6	57.41
7	56.95

To compare this to the percentage of people who watched the videos, we will take an average over each run of the percentage of people that watched the video. To be fully clear in our analysis, we will also split these average out into percentage of the video viewed. Contrasting this with a plot of the above percentages of questions answered correctly, gives us the following plot:



From this plot we can see clearly that in the fourth run of the MOOC course, there was a drop both the

percentage of questions answered correctly and the amount of people who viewed the course videos. We can observe that

Evaluation

Deployment